ACOUSTIC PROFILES IN PROTOTYPICAL VOCAL EXPRESSIONS OF EMOTION

Tom Johnstone, Rainer Banse, and Klaus R. Scherer
Department of Psychology, University of Geneva

ABSTRACT

Acoustic data from a large study on actor portrayals of vocally expressed emotions are reanalysed on the basis of the differences in the accuracy of recognition of the voice samples used. The results show differentiated patterns with respect to the similarity and variability of the mean acoustic profiles for well versus poorly recognized portrayals.

INTRODUCTION

Banse & Scherer [1] report a large-scale study on the expression of emotion in multiple communication modalities, in which 12 professional actors were asked to portray 14 emotions varying in intensity and valence or quality. The results on decoding replicate and extend earlier findings demonstrating the ability of judges to infer vocally expressed emotions with much better than chance accuracy for a large number of emotions. Consistently found differences in the recognizability of different emotions are also replicated. A total of 224 different portrayals, 16 per emotion category, were subjected to digital acoustic analysis to obtain profiles of vocal parameters for different emotions, using a large set of acoustic variables. The data provide first indications that vocal parameters not only index the degree of intensity typical for different emotions but also differentiate valence or quality aspects. In particular, the data are used to test theoretical predictions on vocal patterning based on Scherer's component process model of emotion [2]. While most hypotheses are supported, others need to be revised on the basis of the empirical evidence.

Discriminant analysis and jack-knifing were used to determine how well the 14 emotions can be differentiated on the basis of the vocal parameters measured. The results show remarkably high hit rates and patterns of confusion that closely mirror those found for listener-judges. One of the major results of this study was the identification of typical acoustic profiles for 14 major emotions. However, the portrayals used to compute these profiles varied substantially in the extent to which their emotional content was recognized by listener-judges, despite the fact that they had been preselected for clarity of emotional expression. In this study we report a new, secondary analysis of the earlier data set in order to examine potential differences between acoustic profiles for portrayals that are and that are not well recognized by listener judges. One can argue that portrayals that are well recognized on the basis of vocal expression alone represent prototypical examples of vocal emotion communication. In consequence, their acoustic profiles should represent more closely the acoustic parameters which index different emotional speaker states in natural speech. In contrast poorly recognized portrayals should show greater parameter variation and a less pronounced profile.

METHOD

The mean accuracy of the judgments (computed on the basis of recognition scores ranging from 0 to 12, corresponding to the number of judges who correctly categorized each of the intended emotions as portrayed by the actors) was used to split the vocal utterances into two groups: well recognized vs. poorly recognized (yielding an average score of 8.5 for the well recognized stimuli as compared to 3.2 for the poorly recognized). The respective profiles over 29 acoustic parameters reported previously [1] were computed separately for the two groups of stimuli.

RESULTS

Splitting the utterances produced two groups of utterances for each emotion with substantially different mean recognition scores except in the cases of disgust (difference in means = 2.6) and shame (difference in means = 3.2). These two emotions were badly recognized overall (with overall mean recognition scores of 1.5 and 3.2 respectively) and thus the small difference between the well and poorly recognized groups might be explained as a floor effect. It is likely that the actors found it very difficult to express these emotions and in groping for ways of expressing the requested emotion, either did not produce systematic changes to the vocal signal or consistently produced utterances which were confused with another emotion.

The correlation between the mean profiles for well recognized utterances and those for poorly recognized utterances for each emotion was calculated to provide a measure of profile similarity. The emotions can be divided into three classes; those with low, medium and high correlations between the well versus poorly recognized sample profiles respectively.

The utterances expressing disgust (r=0.02) and interest (r=0.12) fall into the *low correlation* class. As mentioned previously, disgust had a poor overall recognition score. This can be attributed to the lack of any consistent acoustic profile, as shown in Figure 1, and is consistent with previous studies of disgust which show the emotion to be difficult to recognize in speech [3, p.190]. Possibly, the expression of disgust typically involves the use of affect bursts

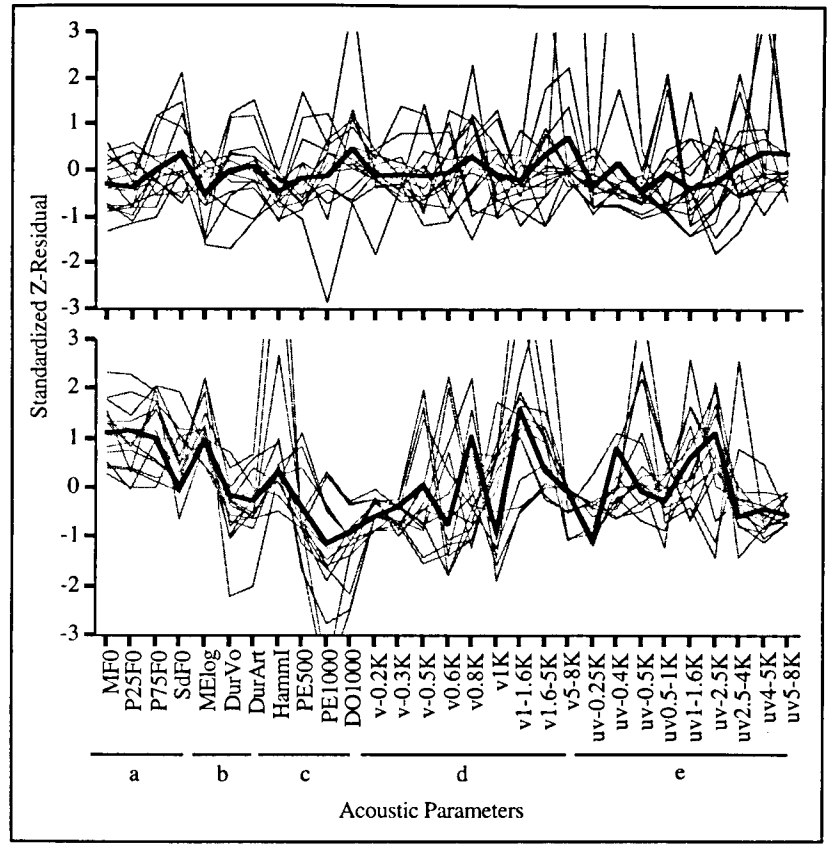


Figure 1. Acoustic profiles of all disgust (top) and hot anger (bottom) utterances, with the mean profiles shown as the dark lines (Acoustic parameters: a) F0 measures b) energy and duration measures, c) high-low frequency ratio, d) voiced spectral parameters, e) unvoiced spectral parameters: see Banse and Scherer [1], Table 6, for a full explanation of the acoustic variables).

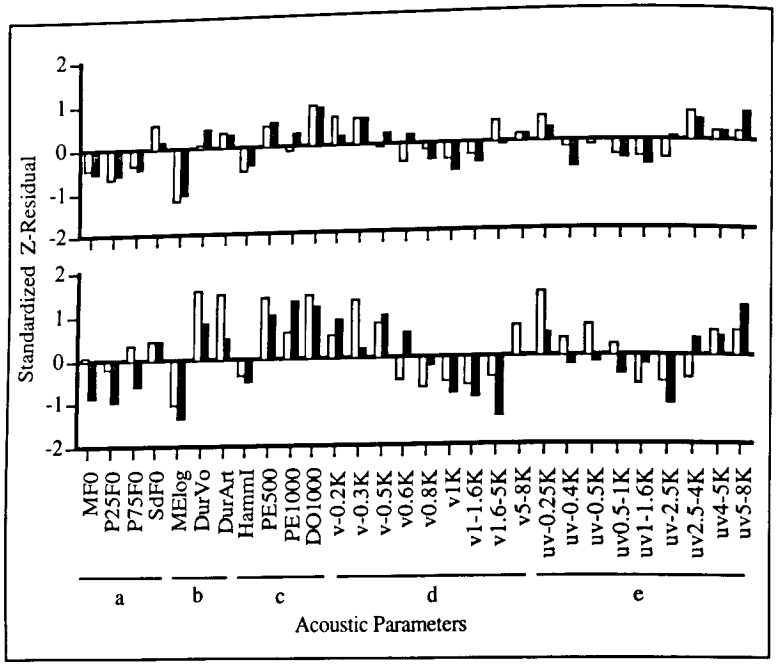


Figure 2. Acoustic profiles of shame (top) and sadness (bottom). White columns are for poorly recognized and shaded columns for well recognized utterances.

rather than the nonverbal modulation of fluent speech [4]. In contrast, interest had a high overall mean recognition score of 11. It is possible that, of the 29 acoustic parameters making up each profile, only a few are used in the expression of interest. Other parameters not measured in the study, such as the type of F0 contour, could play an important part in the expression of interest. Thus the profiles measured in this analysis would not be very well defined despite the high recognition of the utterances.

Utterances expressing the emotions of happiness, cold anger, boredom, pride and panic have *medium sized* correlations between well and poorly recognized group profiles (ranging from r =.37 for pride to .58 for cold anger). These emotions have medium overall recognition scores, implying that the actors were able to express the emotions reasonably well but that there was still considerable variability in the utterances. An examination of profiles indicates that the mean profiles for the poorly recognized utterances are quite similar in shape to those of the well recognized utterances, but usually involve smaller magnitudes. It is possible then that, in these cases, the poorly recognized utterances do not contain sufficient modulation of the relevant acoustic parameters to be identified accurately.

With the exception of shame, all the emotions with *high correlations* between well and poorly recognized utterances had medium to high overall recognition scores. These utterances are generally characterized by well defined acoustic profiles (e.g. the hot anger profile in Figure 1), which would presumably be responsible for the correct recognition of the intended emotion. It is possible that for those emotions which only had medium recognition scores, one or two acoustic parameters which are essential for the expression of the emotion are inconsistently used by the actors. Such idiosyncratic modulation of only a few parameters would not greatly affect the profile correlations. Thus whilst the profiles are consistent and highly correlated, some single important acoustic parameters may vary between actors, leading to poorer recognition of some utterances. It is also possible that in some cases, a number of poorly recognized uterances were not characterized by consistent profiles, due to high variability between speakers. In the cases of sadness and despair, there were significantly higher between-utterance variances for poorly recognized as opposed to well recognized utterances ($t=3.1$, $p<0.05$ and $t=2.7$, $p<0.05$ respectively). Thus the poorly recognized sets of utterances for these emotions did not represent prototypical emotion profiles.

Although utterances expressing shame had well defined profiles, they were very poorly recognized. Comparison of the acoustic profiles of sadness and shame indicates that actors may have been using the sadness prototype when trying to express shame. It is conceivable that, faced with difficulties expressing shame, actors reverted to the more familiar expression of sadness. This is supported not only by the similarity of the profiles for shame and sadness (Figure 2), but also by the large percentage of times shame utterances were falsely categorized as sadness by the judges in the study of Banse and Scherer [1].

CONCLUSIONS

It is apparent that studying actor portrayals of vocal emotion expression can reveal much about the nature of the acoustic parameters involved in the identification of emotion by a listener. At the same time certain emotions either do not seem amenable to consistent portrayal by actors or are not readily recognized by listeners. Certain emotions, such as hot anger and boredom, are portrayed using highly prototypical acoustic profiles which are easily produced by actors and accurately decoded by listeners. Others, while also characterized by quite consistent profiles, are not as well recognized, possibly due to the inability of actors to completely control all the aspects of voice or speech relevant to that emotion. This might be explained in terms of involuntary physiological changes to the vocal apparatus during real emotional episodes, which are inaccessible to voluntary production by actors. Still other emotions, such as interest, although well recognized by listeners, seem be communicated by suprasegmental features other than long term average modulation of the speech signal. Temporal changes in speech parameters such as F0 might be the primary method of encoding such emotions. Finally, disgust would seem to be universally badly encoded and recognized in speech. This could be due to the fact that disgust is more often expressed by brief affect bursts or interjections rather than by modulation of continuous speech.

The secondary analysis of the data set in [1] has shown the utility of using decoding data (i.e. contrasting well versus poorly recognized portrayals) to better understand the role of the encoding of vocally portrayed emotions (as measured by the variation of acoustic profiles). The results of the comparison yield a number of hypotheses which are amenable to further empirical research.

ACKNOWLEDGMENT

The preparation of this paper has been funded by a grant of the Swiss National Fund for Scientific Research (FNRS No. 21-32648.91) in the context of the ESPRIT-BRA VOX workshop programme.

REFERENCES

[1] Banse, R. and Scherer, K. R. (in press), "Acoustic profiles in vocal emotion expression.", *Journal of Personality and Social Psychology.*

[2] Scherer, K. R. (1986), "Vocal affect expression: A review and a model for future research.", *Psychological Bulletin*, vol. 99, pp. 143-165.

[3] Pittam, J., & Scherer, K. R. (1993). "Vocal expression and communication of emotion." In M. Lewis & J. M. Haviland (Eds.), *Handbook of emotions* (pp. 185-198). New York: Guilford Press.

[4] Scherer, K. R. (1994). "Affect bursts." In S. H. M. van Goozen, N. E. van de Poll, & J. A. Sergeant (Eds.), *Emotions: Essays on emotion theory* (pp. 161-196). Hillsdale, NJ: Erlbaum.

TOWARDS ACOUSTIC PROFILES OF PHONATORY QUALITIES

Ailbhe Ní Chasaide and Christer Gobl
Centre for Language and Communication Studies, Trinity College, Dublin, Ireland

ABSTRACT

Voluntary modulations in the mode of phonation constitute an important resource speakers use for the paralinguistic signalling of attitude and emotion. As a first step towards providing profiles of a range of phonatory qualities, this paper presents brief illustrative sketches of some acoustics characteristics associated with modal, tense, breathy/lax, whispery, and creaky voice, as described in [1]. The principal analytic technique on which these illustrations are based was interactive inverse filtering. Source characteristics were measured on the basis of a parametric model of the voice source (the LF model, [2]) fitted to the output of the inverse filter, and from spectral analyses.

INTRODUCTION

Individuals differ in terms of the habitual phonation quality they use, in a way that reflects not only the physical characteristics of their vocal apparatus, but also the linguistic and social group they belong to. As is outlined in the presentation by Laver (this session), speakers also make voluntary short term changes to their mode of phonation as a way of signalling their attitude, mood or emotion. The paralinguistic significance of some phonatory qualities may tend to be universal (e.g., whispery voice tends to give the impression of confidentiality), whereas in other cases, it may be culture or language specific (e.g., creaky voice is associated with bored resignation for some speakers of English).

Our understanding of this aspect of vocal communication is still quite limited. The relative paucity of systematic, quantitative work in this field does not reflect its importance but rather the lack of analytic tools and the many methodological difficulties presented by this kind of research. As a starting point, one needs to be able to make detailed and reliable analyses of the acoustic and physiological correlates of different voice qualities, laryngeal and supralaryngeal. Then there is the difficulty of eliciting appropriate samples of speech. On the one hand, emotionally coloured spontaneous speech would be highly desirable, but on the other hand, reliable analytic comparisons may depend on the speech material being very controlled (more on this in the Conclusions).

This paper presents illustrative sketches of a number of phonatory qualities, based on exploratory work carried out by the authors in recent years [3, 4]. The illustrations are of course tentative, being based on a detailed analysis of just a few utterances spoken with a few of the phonatory qualities that speakers are known to exploit for paralinguistic signalling. The qualities chosen were from the set described by Laver [1], whose descriptions serve as a starting point for our analyses.

DESCRIBING THE SOURCE

The main analysis technique involved inverse filtering of the speech pressure waveform. In order to obtain quantifiable results, a parametric model of differentiated glottal flow (the LF-model, [2]) was matched to the output of the inverse filter. Both the inverse filtering and the matching procedure were carried out for each glottal cycle, using specially designed interactive software allowing optimisation in both the time and frequency domains [5]. From the matched model a number of parameters were subsequently measured. The ones we focused on particularly were EE, RA, RK, RG, OQ and UP. These are explained briefly here, but for a more detailed description, see [7].

EE is the excitation strength and is measured as the negative amplitude of the differentiated flow at the moment of maximum discontinuity. It corresponds to the overall intensity of the signal, so that an increase in EE amplifies all frequency components.

RA is a measure of the return phase (dynamic leakage), which is the residual flow from excitation to complete closure. The acoustic consequence of the return phase is a steeper spectral slope. A large RA corresponds to greater attenuation of the higher frequencies.

RK is a measure of the skew of the glottal pulse: a larger value means a more symmetrical pulse shape. RG is a measure that relates to the duration of the opening branch of the glottal pulse. RK and RG together determine the open quotient, OQ, and they mainly affect the levels of the lower harmonics in the source spectrum. Note that in our definition of OQ, the open phase does not include the interval of the return phase.

UP is the peak glottal airflow, measured only for the oscillatory component of the glottal wave. In our data, UP was calculated indirectly from the other parameters, using a formula suggested by [6].

Aspiration noise is an important source parameter, particularly in breathy and whispery voice. We do not include it in our descriptions, simply because we had no reliable way of measuring it.

Spectral measurements from narrow band spectral sections were also carried out, both on the speech output signal (e.g., Figure 6) and on the source signal, output of the inverse filter (e.g., Figures 2, 3 and 4). Fuller details on these are provided below.

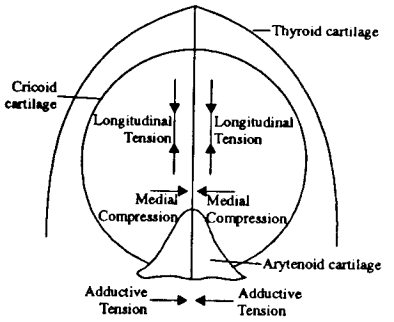


Figure 1. Three laryngeal parameters of muscular tension, after [1].

ACOUSTIC PROFILES

In this section we attempt to illustrate some of the acoustic characteristics of the following phonatory qualities: modal, tense, breathy/lax, whispery and creaky. Although the breathy and lax qualities were separately recorded and analysed, they are rather similar qualities and are treated under a single heading below. Following Laver's descriptions [1] the physiological correlates of these phonatory qualities are presented in terms of three hypothesised dimensions of muscular tension, schematically shown in Figure 1. *Adductive tension* results from contraction of the interarytenoid muscles and is the force which draws the arytenoids together so that the cartilaginous glottis is adducted. *Medial compression* results primarily from contraction of the lateral cricoarytenoid muscle (although the thyroarytenoid muscle can also contribute). It is defined as the force which causes approximation of the vocal processes of the arytenoids so that the ligamental glottis is closed. *Longitudinal tension* is the tension in the vocal folds which results from contraction of the vocalis and cricothyroid muscles, whose primary function is the control of pitch. (For additional descriptions of voice quality see [8, 9].)

The illustrations below of acoustic properties are based on analysis of two

words spoken with the six qualities mentioned above. These words were extracted from recordings of materials read by a male phonetician, who was well practised in the Laver system. The recordings included the Rainbow passage and a number of nonsense words inserted into a carrier frame.

Figures 2, 3 and 4 below illustrate source characteristics in the first vowel of the nonsense word *babber*, which occurred in the frame *Say ---- again*. Figure 2 shows individual spectral sections of the voice source for four phonatory qualities at approximately the midpoint of the vowel. In Figure 3 are shown for these same qualities schematic source spectra, averaged for an interval corresponding approximately to four glottal cycles in the middle of the vowel. These were obtained in the following way. First of all, the spectrum was flattened by adding 6 dB per octave relative to L_0. The spectrum up to 4 kHz was then divided into four frequency bands of 1 kHz. Within each band the average amplitude of the harmonics was calculated and shown relative to L_0. By doing this we should get an idea of how the spectral slope for each of the four qualities deviates from the "ideal" slope of -6 dB per octave in the differentiated glottal flow (the horizontal zero line). In Figure 4 are shown for five qualities the relative levels of the first two harmonics in the source spectrum, measured at the approximate midpoint of the vowel. Figure 5 illustrates the pulse-by-pulse variation in some of the measured source parameters for the first six glottal cycles of the word *strikes*, taken from the Rainbow passage. Finally, Figure 6 illustrates the relative levels of F1 and H1 for approximately the same interval of this word.

Modal

Modal voice is the quality "which phonetic theory assumes takes place in ordinary voicing, when no specific feature is explicitly changed or added (p. 95)" [1]. For this quality, adductive tension, medial compression and longitudinal tension are thought to be moderate, and the ligamental and cartilaginous glottis are thought to vibrate as a single unit. The vocal fold vibration is further described as regularly periodic and efficient, with full glottal closure and thus, without audible glottal frication noise. Recent studies suggest, however, that incomplete glottal closure may be very common even in what is perceived as modal voice [10] and particularly in female speech.

In our analyses, this quality emerged as a relatively efficient mode of phonation, with a fairly strong excitation (EE) and fairly limited dynamic leakage (RA). For an illustration of some source parameter values in the word strikes, see Figure 5. The source spectrum for this quality in the vowel in *babber* exhibited a slope that is slightly greater than the "ideal" description of -12 dB per octave (or -6 dB in the differentiated flow: see Figure 3).

It is important to bear in mind that utterances spoken with modal (or indeed any) quality exhibit considerable dynamic variation as a function of the prosodic and segmental context [7, 11, 12].

Tense

At the laryngeal level, tense voice is thought to involve increased adductive tension and medial compression. The term "pressed phonation" is sometimes used for this quality. A higher degree of tension is likely to be found in the entire production system, and this will have consequences for the respiratory system (a raised subglottal pressure) as well as for the supralaryngeal articulation.

In our measures of tense voice (see, for example, source data in Figure 5) the glottal pulse exhibited a very low dynamic leakage (RA), showing a rather

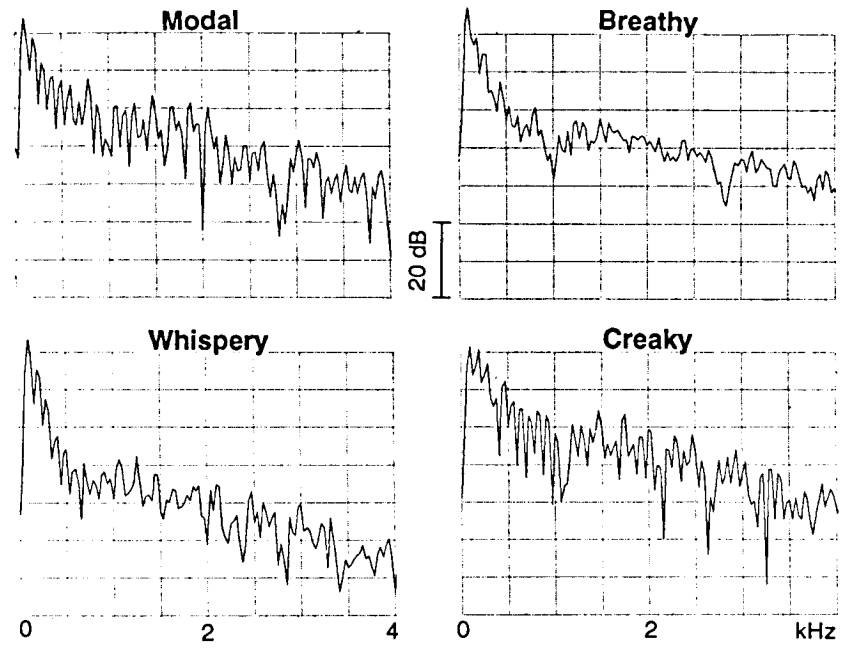


Figure 2. Spectral sections of the source signal at about the midpoint of [æ] *in babber, for four phonatory qualities.*

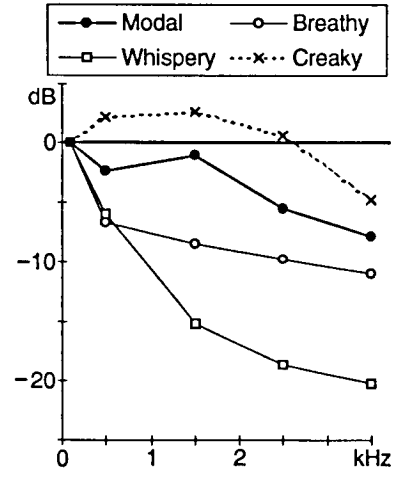


Figure 3. Schematic source spectra in [æ] *of babber, showing within four 1 kHz bands the average deviation from an "ideal" slope (see text).*

instantaneous closure of the vocal folds. The related frequency measure FA was generally higher than for all the other qualities measured in these contexts, and this would imply a flatter spectrum. Relative to modal phonation, the glottal pulse was rather skewed (low RK), the open quotient was lower (low OQ) and RG was higher. The effects of a high RG can be seen in the boosting of H2 relative to H1 for this quality (see illustration for *babber* in Figure 4). Overall, the higher frequencies in the spectrum are relatively dominant. One can get some impression of the relative balance of the lower and higher components of the spectrum from Figure 6 where, for *strikes*, the level of F1 (L_1) is shown relative to that of H1 (L_0). Note the very high L_1 of tense as compared to modal phonation. It is worth noting that for this particular utterance, the excitation (EE) was less

strong for tense than for modal voice (see Figure 5). Thus, the greater intensity of the former in the speech output signal is determined in this instance by these other characteristics of the glottal pulse discussed above.

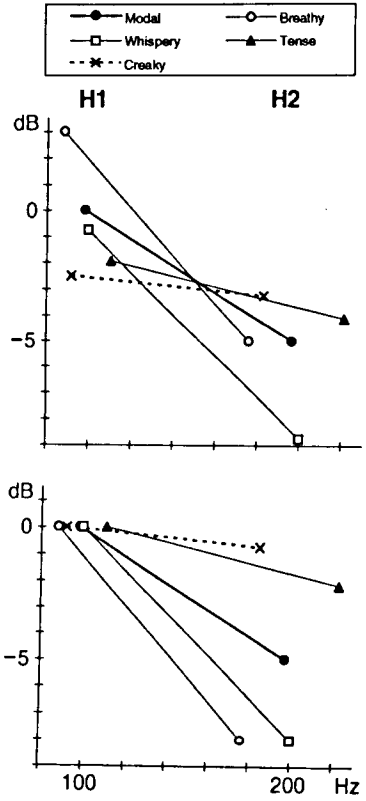


Figure 4. Relative amplitudes of H1 and H2 in source spectra of [æ] *in babber, five qualities. Values shown in absolute terms (upper panel) and normalised to* L_0 *of modal voice (lower panel).*

Breathy/Lax

Breathy voice is described as having minimal adductive tension, low longitudinal tension and weak medial compression, with the result that the vocal folds never come fully together and generate audible frication noise. At the laryngeal level, lax voice is described as being rather similar to breathy voice. It may differ in the extent to which laryngeal tensions are reduced: Laver suggests that lax voice may be slightly closer to modal than breathy voice. It is further postulated as having a lesser degree of tension in the entire speech apparatus.

As expected, the glottal pulse for both breathy and lax voice had rather high dynamic leakage (RA): see source values for lax voice in Figure 5. The related frequency measure FA is lower than for modal voice. The glottal pulse also has a high open quotient (OQ) with a long opening branch (low RG). These last characteristics contribute to the relative boosting of H1, a frequently observed spectral characteristic. See for example, the relatively strong H1 for breathy voice in the upper panel of Figure 4, and the rather low values for L_1 relative to L_0 for lax voice in Figure 6.

Whispery

At the laryngeal level whispery voice is thought to be characterised by low adductive tension and moderate longitudinal tension. The degree of medial compression may vary, and with it the size of the triangular opening of the cartilaginous glottis. With weak whisper medial compression is moderate and the opening may include part of the ligamental glottis. Whisper with a higher intensity is thought to have higher medial compression and a smaller opening of the cartilaginous glottis. It is suggested that laryngeal vibration is confined to the portion of the ligamental glottis which is adducted, and the whispery component to the opening between the arytenoids.

Whispery and breathy voice may form an auditory continuum. Although there may be no clear border line [1], they may nevertheless be auditorily distinguished by the relative dominance of the periodic and noise components: the noise component would be relatively greater for whispery voice, whereas the periodic component would be dominant in breathy voice.

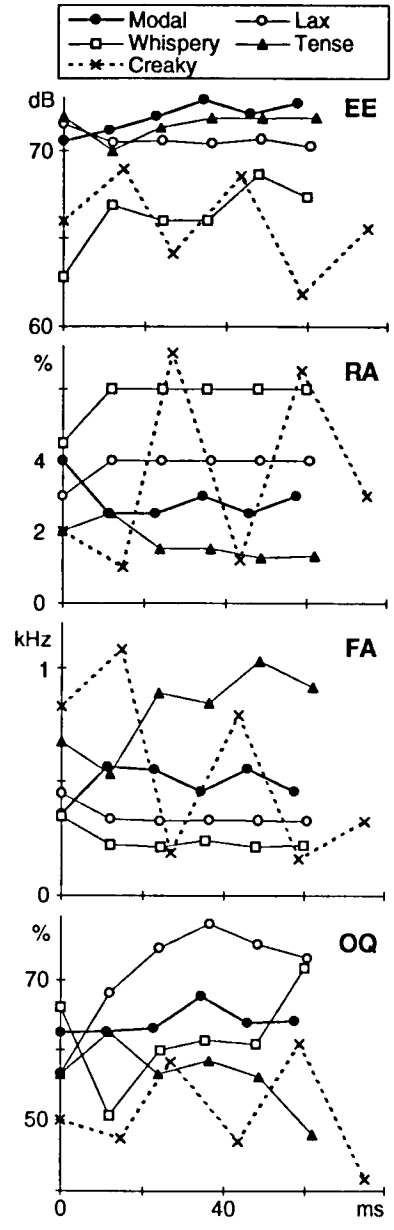


Figure 5. Source data for EE, RA, FA and OQ in the first six glottal cycles of strikes, for five phonatory qualities.

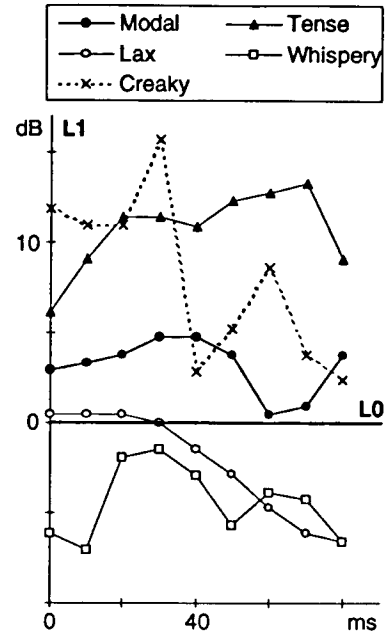


Figure 6. The levels of F1 (L_1*) relative to H1 (*L_0 *= 0 dB) for the first 80 ms in strikes for five phonatory qualities.*

In our measurements, the glottal pulse for whispery voice was found to differ from that of breathy voice in the considerably weaker excitation (EE), the higher dynamic leakage (RA), highest of all measured samples, and in the smaller open quotient (OQ). Calculated peak airflow values (UP) were also considerably lower. As a correlate of these last two factors, one would expect H1 to be less boosted than for breathy voice, and this can be seen, for example, in the upper panel of Figure 4, where the levels of H1 and H2 are shown in absolute terms for *babber*. As can be seen in the lower panel of this figure, however, the relative levels of H1 and H2 are very similar, presumably because of the steeper spectral slope of whispery voice.

This last can be seen in Figure 3, and is a clear consequence of the very high RA and low FA values.

Creaky

Creaky voice, being a mix of creak and voicing, is likely to involve the high adductive tension, high medial compression and low longitudinal tension, characteristic of creak. In the production of creak, the folds are thought to be relatively thick and compressed. The ventricular folds may also be somewhat adducted, so that their inferior surfaces come in contact with the superior surfaces of the true vocal folds, resulting in a rather thick vibrating structure. Because of the high adductive tension only the ligamental glottis is thought to vibrate. The f_0 and amplitude of consecutive glottal pulses are described to be irregular and the airflow rate has been observed to be very low [8].

In our recordings of creaky voice, the occurrence of creak was intermittent, in the sense of irregularity in successive glottal pulses. It did not occur in the word *babber* but did in the word *strikes*. In the latter, there is a clear alternation of two very different types of glottal pulse (see Figure 5). One is efficient, with a reasonably strong excitation (though not as strong as for modal), very low dynamic leakage (RA) and consequently a high FA. The other pulse is inefficient, with a very weak excitation, very high dynamic leakage (and a low FA). These two types of pulses should have very different source spectra. Both pulses show a relatively high degree of skew (low RK), a low open quotient and a high RG. The short open phase and the low calculated UP which are also found here are consistent with the low airflow rate reported for this quality and should affect the lower end of the spectrum by reducing its level. Note in Figure 6, that despite fluctuations, L_1 dominates L_0. Given the long closed phase, there is less damping and this could also contribute to the strong ringing of F1.

In the *babber* utterance successive glottal pulses were not irregular and resembled more the strong efficient type of pulse described above. Note the strong H2 and rather weak H1 for this quality in Figure 4 and the relative dominance of higher frequencies in Figure 3. In many ways, the spectral characteristics of this quality resembled those of tense voice.

CONCLUSIONS

We would emphasise that these illustrations are tentative and should not be taken as definitive accounts but rather as a first step toward more comprehensive profiles of different phonatory qualities. This is not only because they are based on a very limited number of utterances. Other factors need to be borne in mind when it comes to interpreting these kinds of source data.

First of all, cross-speaker variation can be quite large. For example, in the important source parameter RA, we have observed cross-speaker differences as great or greater than the differences shown here for a single speaker who has intentionally varied his voice quality. We feel therefore, that the relevant measures of phonatory quality may need to be expressed in terms of deviations from a given speaker's baseline values, rather than in absolute terms. A similar point has been made by other researchers, e.g., [13] as a result of trying to characterise linguistically distinctive phonatory qualities.

Another factor, alluded to briefly before, concerns the considerable dynamic variation that can occur within a single voice quality. These variations appear to be (at least sometimes) conditioned by the prosodic and segmental context and have been illustrated in earlier work [7, 11, 12, 14]. This consideration seriously constrains the kinds of speech materials that can usefully be compared using these analysis techniques. Furthermore, profiles of individual phonatory qualities can not be static, but will need to take account of this dynamic modulation.

Clearly, there is much work to be done before one will have available comprehensive descriptions of the acoustic correlates of particular phonatory qualities. Yet even partial descriptions may provide useful reference material for looking at more spontaneously occurring speech, where the mode of phonation has been varied for paralinguistic purposes. Furthermore, such descriptions should eventually provide a basis for resynthesis, which should allow one to explore directly the paralinguistic colouring associated with individual voice qualities.

ACKNOWLEDGEMENTS

This work was supported by the Esprit- BRA Working Group, no. 6975, VOX. We are also grateful to Francis Nolan who was our informant.

REFERENCES

[1] Laver, J. (1980), *The phonetic description of voice quality*, Cambridge: Cambridge University Press.

[2] Fant, G., Liljencrants, J. and Q. Lin (1985), "A four-parameter model of glottal flow", *STL-QPSR*, Vol. 4/1985, pp. 1-13.

[3] Gobl, C. (1989), "A preliminary study of acoustic voice quality correlates", *STL-QPSR*, Vol. 4/1989, pp. 9-22.

[4] Gobl, C. and Ní Chasaide, A. (1992), "Acoustic characteristics of voice quality", *Speech Communication*, 11, pp. 481-490.

[5] Ní Chasaide, A., Gobl, C. and Monahan, P. (1992), "A technique for analysing voice quality in pathological and normal speech", *Journal of Clinical Speech & Language Studies*, Vol. 2, pp. 1-16.

[6] Fant, G and Lin, Q. (1988), "Frequency domain interpretation and derivation of glottal flow parameters", *STL-QPSR*, Vol. 2-3/1988, pp. 1-21.

[7] Ní Chasaide, A. and Gobl, C. (1993), "Contextual variation of the vowel voice source as a function of adjacent consonants", *Language and Speech*, 36, pp. 303-330.

[8] Catford, J.C. (1964), "Phonation types: the classification of some laryngeal components of speech production", in D. Abercrombie, D.B. Fry, P.A.D. MacCarthy, N.C. Scott, and J.L.M. Trim (eds.), *In Honour of Daniel Jones* (pp. 26-37), London: Longmans.

[9] Ladefoged, P. (1971), *Preliminaries to linguistic phonetics*, Chicago: The University of Chicago Press.

[10] Södersten, M. (1994), "Vocal fold closure during phonation", Ph.D. thesis, Studies in Logopedics and Phoniatrics No. 3, Huddinge University Hospital, Stockholm.

[11] Gobl, C. (1988), "Voice source dynamics in connected speech", *STL-QPSR*, Vol. 1/1988, pp. 123-159.

[12] Pierrehumbert J.B. (1989). A preliminary study of the consequences of intonation for the voice source. *STL-QPSR*, Vol. 4/1989, pp. 23-36.

[13] Traill, A. and Jackson, M. (1987), "Speaker variation and phonation types in Tsonga nasals", *UCLA Working Papers in Phonetics*, 67, pp. 1-28.

[14] Gobl, C., Ní Chasaide, A. and Monahan, P. (1995), "Intrinsic voice source characteristics of selected consonants", *Proc. of the XIIIth ICPhS*, Stockholm.

PROSODIC ANALYSIS OF BABBLING AND FIRST WORDS: A COMPARISON OF ENGLISH AND FRENCH

Marilyn May Vihman (Southeastern Louisiana University), Barbara L. Davis (University of Texas at Austin) and Rory DePaolis (Southeastern Louisiana University)

ABSTRACT

Prosodic analysis was undertaken for disyllabic vocalizations sampled during the transition to language in four children, two acquiring English and two French. Early mastery was predicted for prosodic parameters whose natural manifestations are supported by the stress system of the ambient language (final lengthening in French, joint increase in pitch and amplitude on the first syllable in English). Results supported the model for stress acquisition in French but not in English .

INTRODUCTION

The present study is part of a long-term effort to trace the child's course from phonetic mastery, or emergent speech motor control, to phonological system in the transition to language. By comparing profiles over time for children acquiring different languages it is possible to begin to distinguish among the relevant phonetic abilities, identifying (1) those which are available from the beginning of speech production, (2) those which emerge under the specific influence of the ambient language, and (3) those which are acquired still more gradually, with wide variation at first across individual children exposed to the same language [1]. Although the acquisition of segmental, syllabic and prosodic patterns necessarily involves a complex interaction between these different abilities, recent work has begun to map out some developmental aspects of the phonetic "territory" as far as segmental patterns are concerned. For example, early influence of the ambient language has been found in the prelinguistic period in characteristics of the vowel space [2] and in the proportion of labials produced (at about 10 months [3]), despite the fact that in the languages in question (Arabic, Cantonese, English and French for vowel space, English, French, Japanese and Swedish for labials) the children's productions generally agree in showing low front and central vowels and front consonants (labials and dental/alveolars).

There is reason to expect prosodic parameters - essentially, variation in pitch, duration, and loudness - to be particularly salient to children and thus to reflect ambient language influence particularly early. The idea, advanced by Lewis among others [4], that both the physical dimensions and the affective content of the prosody of infant-directed speech attracts the child's attention long before he or she has begun to recognize segmental patterns has received strong support from experimental studies of perception over the past decade [5, 6]. Furthermore, advances in the discrimination of ambient vs. foreign language and in differential response to normally segmented vs. interrupted speech units have been shown to be based almost entirely on prosody within the first six months of life [7]. Only in the latter part of the first year, after the child has begun to produce canonical babbling [8] and is thus equipped with a rhythmic motor **frame** for speech [9], do we find evidence of comparable perceptual advances based on the **phonetic content** of the speech stream. From the point of view of production, prosody characterizes whole units (sequences of syllables, words, phrases), and thus might be expected to be accessible for accurate replication earlier than individual consonant-vowel sequences, since early phonology is thought to be based on holistic units, not individual segments [10, 11, 12]. Finally, an understanding of the development of prosody is of considerable importance for phonological development insofar as the dual role of rhythm as a regulator of

motor behavior in general and of speech production in particular (the constraints of each language defining the particular manifestations of the requirement of rhythmicity in speech) may be seen to constitute an essential link between biological and linguistic structure [13].

However, prosodic analysis - especially with regard to stress-accent systems - presents challenges not found in the analysis of segmental patterns. Although amplitude, pitch and duration are in principle individually controllable components of the adult production system, they are tightly intercorrelated in stressed syllables - and their developmental patterns need not be the same. It is thus not surprising that while cross-linguistic prosodic analyses have indeed suggested early influence of the ambient language [14, 15, 16, 17], few longitudinal studies have been published to date. The present study provides preliminary results from a planned longitudinal comparison encompassing the transition from prelinguistic vocalizations to a vocabulary of 50 words or more in four languages, with five to ten subjects in each language group.

Finally, since the percept of stress - a linguistic, not a physical parameter - is multiply determined and is not a "natural" component of prelinguistic production, two quite different approaches are possible: The "top-down" approach begins by evaluating infant vocalizations for the presence of something that sounds like stress; acoustic analysis is then undertaken to attempt to ascertain the physical base for that percept [18, 19]. The "bottom-up" approach, which we will follow here, avoids adult perceptual judgments of stress in the infants' productions and focusses acoustic analysis on differences between the syllables of a disyllabic vocalization [cf . also 20].

The prosodic systems of English and French

The stress systems of English vs. French present sharp and measurable contrasts. In the lexicon used with one-year-olds, English has a predominantly trochaic rhythm [21]. That is, the concomitants of stress - considerably higher amplitude, pitch change, and a full (i.e., not reduced) and somewhat longer vowel [22] - together characterize the first ("stressed") syllable of most of the two-syllable words the children hear. In French, on the other hand, stress is phrase-final in the adult system, meaning that early words are necessarily iambs, and the stressed syllable is characterized by pitch change and vowel lengthening but a **decrease** in amplitude [23]. In addition, a relatively higher percentage of French phrases end in rising pitch in comparison with English [24]. Finally, although the adult models for the early words of children learning English typically bear initial stress, they are also characterized by a degree of final-syllable lengthening [23].

The problems posed for the child learner by the two contrasting prosodic systems are correspondingly different. Let us suppose that there is an inherent tendency to final syllable lengthening (as suggested by Laufer [25]), but that the placement of relatively greater amplitude or of the major pitch change point of a unifying prosodic envelope is free to vary, depending on the child's whim - or perhaps on degree of sensitivity to the adult model. Let us assume further (as suggested by Allen [26]) that the natural physiological tendency is to increase amplitude and pitch together. Early accommodation to the stress pattern of French, then, would involve (increasingly with lexical development, as the child "enters into the adult language"):

- exaggeration of the natural tendency to **final syllable lengthening**;
- placement of **pitch change on the final syllable**;
- use of a relatively **high proportion of rising pitch contours**;

With regard to amplitude, **no developmental trend** is expected, since the natural tendency to increase amplitude with pitch is in conflict with the phonetic manifestation of stress in the adult model (decrease in amplitude on the last syllable).

For children acquiring English, on the other hand, we expect no developmental trend with regard to duration, in view of the conflict between lengthening as a concomitant of stress (on the initial syllable) and as a feature of phrase-final position; we do expect (increasingly, with lexical development):

- placement of **pitch change on the initial syllable**;
- use of a relatively **smaller proportion of rising contours**;
- **higher amplitude on the first syllable**.

In summary, then, we are testing the following broad hypotheses:

(1) Cross-linguistic differences in several prosodic parameters will be apparent by the end of the study, when the children have acquired a sizable lexicon, but not by the first two sessions, which are largely prelinguistic;

(2) A developmental trend toward a significant difference between first and second syllable will be evident by the end of the study for **amplitude** in English and for **duration** in French;

(3) A significant increase in **mean Fo** will mark the **second** syllable in French as more rising contours are produced, while a smaller increase in Fo will mark the **first** syllable in English, as falling contours coontinue to predominate. In addition, **pitch range** will increase on the second syllable in French, due to the occurrence of both rising and falling contours.

SUBJECTS AND DATA

Data from four children were selected from two groups followed longitudinally as they acquired English and French. All of the children were audiorecorded on a weekly (Austin, Texas) or biweekly (Paris, France) basis in spontaneous play sessions in the home over the period of transition from babbling to words; each of the children wore a small microphone hidden in the pocket of a cloth vest.

Four sessions were identified for analysis on the basis of lexical advance: A 0-word session, in which fewer than 4 different recognizable word types were produced spontaneously, and 4-word, 15-word and 25-word sessions, roughly corresponding to a cumulative vocabulary of 10, 30 and 50 words [27]. The children's age ranges were 9-18 months (English) and 9-17 months (French). Twenty disyllablic vocalizations (including both words and babble) were selected for analysis at each of the four word points for each infant.

ACOUSTIC ANALYSIS

Disyllables were extracted from the audio recordings and digitally recorded. Commercially available software (Computerized Speech Laboratory, Soundscope) was used to measure the three main correlates of stress: fundamental frequency, intensity and duration. Fundamental frequency was based on inspection of the narrow band spectogram as well as on automatic pitch contour analysis (peak-picking and autocorrelation). Mean and peak intensity were derived from the vowel segment of the waveform and computed by the software programs. The extent of the syllable rime was estimated from the wide band spectogram. Rime initiation was taken to be at the onset of the first broad spectrum glottal pulse, termination at the point of marked decrease in higher formant energy. In the case of transition to or from glides and liquids, the transition was divided between consonant and vowel segments. Rime duration was then measured automatically.

RESULTS

The small sample size - two subjects in each language group - affords little power for formal statistical analysis. Nevertheless, we performed a 2 (languages) by 2 (subjects) by 4 (word point) analysis of variance on four measures (amplitude and duration for first to second syllable-rime ratios and mean Fo for first and second syllable, with "language" and "word point" as fixed effects. The only significant result was a language-group difference in second-syllable pitch ($p<.05$, one-tailed). Other apparent effects (language group effects for amplitude ratio and first-syllable pitch) failed to attain statistical significance.

Direct inspection of the measurements obtained does permit us to draw tentative conclusions regarding the nature and degree of ambient language influence on early prosodic development. We will take up each of the prosodic variables in turn.

Duration

Figure 1 displays mean duration of the first and second syllable rimes for all four subjects. Despite differences in absolute duration (note the extremely high values for N), the relationship between the length of the two syllable rimes for each child reveals a consistent within-group effect: The French children (Laurent and Charles) differ at the 0-word point but are closely alike in the last two sessions, in which final syllable lengthening is clearly established. Figure 2 displays the first- to second-rime ratio for the French children, revealing convergence on a ratio of about 1:1.6, which corresponds closely to earlier accounts for both adults [23, 28] and children [26]. On the other hand, the two American children show contrasting trends: Final syllable lengthening decreases steadily across sessions for Nico (to 1:.98 at the last session) but increases from the 4-word point on for Cameron (to 1:1.6).

Amplitude

The two French children show contrary developmental trends in the first-to-second syllable ratio for intensity (Figure 3), with one of the children (Charles) showing higher amplitude on the second syllable at the 4- and 25-word points only. Contrary to our expectations, the American children agree in maintaining relatively even amplitude for the two syllables throughout the period analyzed.

Pitch

Two measures of pitch were taken: We analyzed mean and range of Fo in each syllable rime as an indirect indication of the child's placement of the **major pitch change**. Figure 4 displays the results for mean Fo. For one subject in each language group (right-side panels), the two syllables differ at the 25-word point, while for the remaining two subjects there is little difference. The French child Charles maintains higher mean pitch on the second syllable from the 4-word-point on, reflecting early and relatively consistent use of a dominant rising pattern. The American child Cameron, on the other hand, shows slightly higher pitch on the **first** syllable in most sessions.

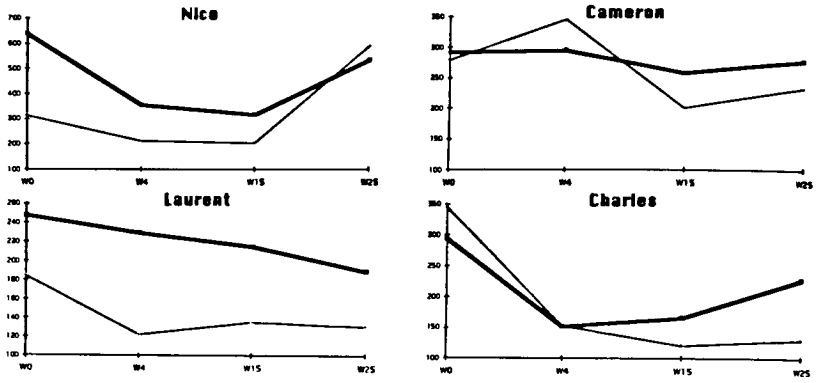


Figure 1. Mean duration of first (fine line) vs. second syllable rime. Note: Scales (in msecs) differ by subject.

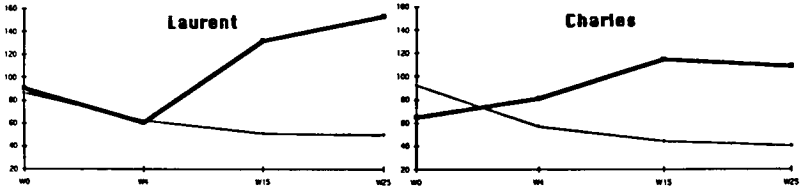


Figure 2. Duration ratio of second to first syllable rime for French subjects.

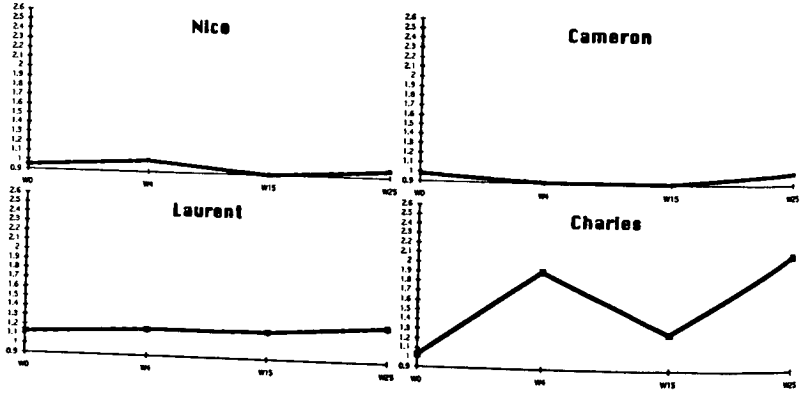


Figure 3. Amplitude ratio (in rms) of second to first syllable.

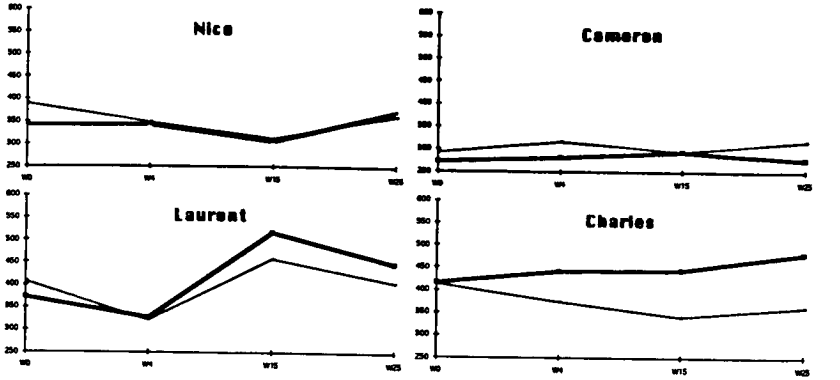


Figure 4. Mean Fo (in Hz) for first (fine line) vs. second syllable.

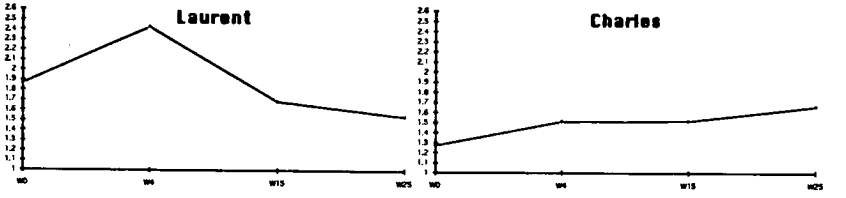


Figure 5. Pitch range of first (fine line) vs. second syllable for French subjects.

The results for range are striking for the French children (Figure 5). What is represented here is the range of Fo in the first compared with the second syllable. The two children show a remarkable degree of agreement from the 15-word-point on, when the second syllable shows a pitch range which is more than 70 herz greater than the range of the first syllable.

DISCUSSION

In our introductory remarks we outlined a set of specific predictions as to the outcome of our prosodic analyses, deriving these predictions from an implicit acquisition model. This model assumes that the child (1) draws on phonetic (speech production) resources which become available over the course of the first year as a result of neurophysiological maturation as well as motoric "exercise" (babbling) in an interactive, affectional social context, (2) is guided by increasing familiarity with and attunement to the segmental and prosodic patterns of the ambient language, and (3) is driven by emerging representational (cognitive) abilities to systematize those patterns, developing an internally consistent set of word production templates out of the "vocal motor schemes" [29], or syllabic frames with increasingly varied phonetic content [9], developed through babbling. Earlier work has shown that phonological systematicity begins to emerge after the first ten or more words have been produced and is identifiable in the child's idiosyncratic patterns by what we term the 15- or 25-word point [1]. How do the findings of the present study fit into this framework?

Children can be expected to acquire most readily those apects of the ambient prosodic system which correspond to physiologically natural tendencies. The predictions of this model are borne out rather well in the French data: Final syllable lengthening - a tendency found within the first year which is also supported by the French stress system - is

established at a comparable level for the two subjects as they increase their lexical store. On the other hand, the rising intonation which characterizes much of French speech - but which requires voluntary effort to counteract the natural tendency to a falling pitch in the course of vocalization [30] - is manifested by only one of the two children (Charles). Interestingly, Charles is the only one of the four children to show higher amplitude on one syllable; in his case, increase in pitch is matched (at least at the 25-word point) by increase in intensity.

The findings for the American children suggest that the English trochaic strong-weak pattern is not easily mastered, despite its perceptual familiarity from an early age [21]. Our finding of inter-subject variability in final syllable lengthening agrees with some earlier studies (e.g., Snow [18], whose first session corresponds developmentally to our last) and can be taken to reflect the conflict mentioned earlier between natural tendency and ambient linguistic model. Both pitch and amplitude are relatively even across the two syllables of our subjects' vocalizations - a finding reported in other studies of English stress acquisition by somewhat older children [19, 31], in which "overarticulation" has been found to characterize the production of unfamiliar word patterns at least to age two years.

ACKNOWLEDGEMENTS

We would like to express our thanks to Edwin Kay (Lehigh University) for carrying out the statistical analysis and to Christine Matyear (University of Texas at Austin) for her help with the acoustic analysis. This work was supported in part by NICHD-R01-HD27733-03.

REFERENCES

[1] Vihman, M. M. & Boysson-Bardies, B. de (1994), "The nature and origins of ambient language influence on infant vocal production and early words", *Phonetica*, vol. 51, pp.159-169.

[2] Boysson-Bardies, B. de, Hallé, P. A., Sagart, L. & Durand, C. (1989), "A crosslinguistic investigation of vowel formants in babbling", *Journal of Child Language*, vol. 8, pp. 511-524.

[3] Boysson-Bardies, B. de & Vihman, M. M. (1991), "Adaptation to language: Evidence from babbling and first words in four languages", *Language*, vol. 67, pp. 297-319.

[4] Lewis, M. M. (1936), *Infant speech*, New York: Arno Press.

[5] Fernald, A. (1984), "The perceptual and affective salience of mothers' speech to infants", in L. Feagans, C. Garvey & R. Golinkoff (eds.), *The origins and growth of communication*, Norwood, NJ: Ablex.

[6] Fernald, A. (1991), "Prosody in speech to children: Prelinguistic and linguistic functions", in R. Vasta (ed.), *Annals of Child Development*, vol. 8, London: Jessica Kingsley.

[7] Jusczyk, P. W. (1994), "Infant speech perception and the development of the mental lexicon", in J. C. Goodman & H. C. Nusbaum (eds.), *The development of speech perception*, Cambridge, MA: MIT Press.

[8] Oller, D. K. (1980), "The emergence of the sounds of speech in infancy", in G. Yeni-komshian, J. F. Kavanagh & C. A. Ferguson (eds.), *Child phonology, I: Production*, New York: Academic Press.

[9] MacNeilage, P. F. & Davis, B. L (1990), "Acquisition of speech production: Frames, then content", in M. Jeannerod (ed.), *Attention and performance, XIII*, Hillsdale, NJ: Lawrence Erlbaum.

[10] Ferguson, C. A. & Farwell, C. B. (1975), "Words and sounds in early language acquisition", *Language*, vol. 51, pp. 419-439.

[11] Macken, M. A. (1979), "Developmental reorganization of phonology", *Lingua*, vol. 49, pp. 11-49.

[12] Vihman, M. M. & Velleman, S. L. (1989), "Phonological reorganization", *Language and Speech*, vol. 32, pp. 149-170.

[13] Kent, R. D. , Mitchell, P. R. & Sancier, M. (1991), "Evidence and role of rhythmic organization in early vocal development in human infants", in J. Fagard & P. H. Wolff (eds.), *The development of timing control and temporal organization in coordinated action*, Oxford: Elsevier Science.

[14] Bacri, N., Boysson-Bardies, B. de & Hallé, P. A. (1989), "Prosodic processing in French and American infants' babbling", *Proceedings of Speech Research '89*, pp. 5-8, Budapest: Linguistic Institute of the Hungarian Academy of Sciences.

[15] Levitt, A. G. & Wang, Q. (1991), "Evidence for language-specific rhythmic influences in the reduplicative babbling of French- and English-learning infants, *Language and Speech*, vol. 34, pp. 235-249.

[16] Whalen, D. H., Levitt, A. G. & Wang, Q. (1991), "Intonational differences between the reduplicative babbling of French- and English-learning infants", *Journal of Child Language*, vol. 21, pp. 59-83.

[17] Hallé, P. A., Boysson-Bardies, B. de & Vihman, M. M. (1991), "Beginnings of prosodic organization: Intonation and duration patterns of disyllables produced by Japanese and French infants", *Language and Speech*, vol. 34, pp. 299-318.

[18] Snow, D. (1994), "Phrase-final syllable lengthening and intonation in early child speech", *Journal of Speech and Hearing Research*, vol. 37, pp. 831-840.

[19] Pollock, K. E., Brammer, D. M. & Hageman, C.F. (1993), "An acoustic analysis of young children's productions of word stress", *Journal of Phonetics*, vol. 21, pp. 183-203..

[20] Robb, M. P. & Saxman, J. H. (1990), "Syllable durations of preword and early word vocalizations", *Journal of Speech and Hearing Research*, vol. 33, pp. 583-593.

[21] Jusczyk, P. W., Cutler, A. & Redanz, N. J. (1993), "Infants' preference for the predominant stress patterns of English words", *Child Development*, vol. 64, pp. 675-687.

[22] Lehiste, I. (1970), *Suprasegmentals*, Cambridge, MA: MIT Press.

[23] Delattre, P. C. (1966), "A comparison of syllable length conditioning among languages", *International Journal of Applied Linguistics*, vol. 4, pp. 182-198.

[24] Rossi, M. (1980), Prosodical aspects of speech productions, *Travaux de l'Institut de Phonétique d'Aix*, vol. 6, pp. 49-72.

[25] Laufer, M. Z. (1980), "Temporal regularity in prespeech", in T. Murry & A. Murry (eds.), *Infant communication*, Houston, TX: College Hill Press.

[26] Allen, G. D. (1983), "Some suprasegmental contours in French two-year-old children's speech", *Phonetica*, vol. 40, pp. 269-292.

[27] Vihman, M. M. & Miller, R. (1988), "Words and babble at the threshold of lexical acquisition", in M. D. Smith & J. L. Locke (eds.), *The emergent lexicon*, New York: Academic Press.

[28] Benguerel, A.-P. (1971), "Duration of French vowels in unemphatic stress", *Language and Speech*, vol. 14, pp. 383-391.

[29] Vihman, M. M., Velleman, S.L. & McCune, L. (1994), "How abstract is child phonology?", in M. Yavas (ed.), *First and second language phonology*, San Diego: Singular Press.

[30] Kent, R. D. & Murray, A.D. (1982), "Aoustic features of infant vocalic utterances at 3, 6, and 9 months", *Journal of the Acoustic Society*, vol. 72, pp. 353-363.

[31] Allen, G. D. & Hawkins, S. (1978), "The development of phonological rhythm", in A. Bell & J. B. Hooper (eds.), *Syllables and segments*, Amsterdam: North-Holland Publishing Company.

A DEVELOPMENTAL MODEL OF ACQUISITION OF RHYTHMIC PATTERNS: RESULTS FROM A CROSS-LINGUISTIC STUDY.

Gabrielle KONOPCZYNSKI
Laboratoire de Phonétique, Université de Franche-Comté
F 25030 BESANÇON-CEDEX

ABSTRACT

The acquisition of rhythm has been overlooked in the study of child language. As the phonological rhythm is quite different from one language to the other, a cross-linguistic study of acquisition of rhythm could provide good insights about universals and language-specific aspects of vocal production. We shall study the acquisition of rhythmical patterns in three groups of languages contrasting by their rhythm: syllable-timed languages (French, Hungarian), stress-timed languages (English, Portuguese) and in-between languages (Spanish). The final aim is to arrive at a predictive developmental model, able to predict the emergence of the correct rhythm of a given language and the errors a child will make in acquiring the stress pattern of the target language.

INTRODUCTION.

In recent years, there has been growing interest in cross-cultural studies of child development. However, very few cross-linguistic studies in early language have been conducted and even fewer are found for the acquisition of prosody. This is rather surprising as comparative studies in this domain could provide good insights about universal and language specific aspects of vocal production.

My aims in conducting cross linguistic studies are threefold: 1) Predict emergence of the stress constraints and stages of acquisition of rhythm in different languages, knowing their rhythmical pattern in adult language, 2) Predict errors in acquiring the stress pattern of a given language, 3) Emphasize the role of language specific factors guiding the child towards a solution unique to each language, either within a particular typological set of related languages, or within a framework of possible universal acquisition processes.

The model I shall propose aims for descriptive adequacy, but the goal is to arrive at a valid explicative model, and finally at a predictive developmental model. My current model, which follows a preliminary model (2a,b) focuses on the acquisition of phrasing, which, in the limits of the utterance, can be called phonological rhythm.

To begin, I will define what I call "phonological rhythm" and, next fix the limits of what I call the "emergent language", and "turning point, or pivotal" period. I will then give information about the rhythmical pattern of each language studied, the population, data and results.

The concept of phonological rhythm is not always clear. It is generally said that

> "rhythm is the structure of a sequence , i.e. the relationship or set of relationships among the units making up that structure. This definition leaves open what those units are; they can be features, segments, syllables, words, phrases, etc." (Allen 1980: 227) [1]

As reference units, I chose the duration of the whole utterance on the one hand, and the duration of the syllables on the other hand, because the syllable seems to be a unit of both perception and a production, and it is less sensitive to variations in flow than the vowel.

Rhythm is called 'phonological' because it explicitly deals with the temporal-sequential constraints of a specific language, and it can be best understood

within the framework of this linguistic level. For most languages, the basic rhythm, which is the rhythm of neutral utterances, is mainly determined by the structure of the syllable and the organization of pauses and accents. From this point of view, languages are divided into different groups, typologically related, although they may have different origins. Thus, I chose two languages in each group: Languages which are said to show either a tendency towards syllable-timing or whose stress has a strictly demarcative function, indicating either the end (French) or the beginning of an utterance (Hungarian); Languages which are said to have a tendency towards stress-timing, as English, which is the typical stress timed language, and to a lesser extent Portuguese spoken in Portugal, and finally, languages which are in-between, as Spanish. This is an oversimplified classification of these groups of languages. In everyday spontaneous speech, a neutral utterance seldom appears. However, for emergent language, these descriptions have the advantage of being efficient and easy to apply to child language which is not yet very complex in syntax or semantics.

Finally, one last point has to be specified before I turn to the topic of the child's rhythm. In previous studies [2] I demonstrated that during the pivotal period it is striking that the child already knows how to use differentiated utterances appropriate to the context. When an auditory analysis was compared with an analysis of the situational context, there was a positive correlation between the utterance context, type of utterance, auditory and acoustic characteristics of the utter-ances. Thus, it appears that babbling is neither egocentric nor monolithic. On the contrary, it contains various types of utterances. For example, when the baby was alone, he emitted non communicative vocalizations to which listeners were unable to attribute meaning. However, in an interactive situation, the sound production was more stable and a majority of listeners were able to classify the utterances into categories such as questions, statements, callings, etc. These were called Proto- or Pseudo-Language (PL). These results were reduplicated in different experimental situations, and are consistent with others. In this paper, I shall deal only with PL emitted in interaction.

The study focuses on the child at the "turning point"' or "pivotal period" from 8; or 9; on, when s/he is still in a prelinguistic stage, to 24; when s/he has already acquired many words and combinations of these. This stage is much more than just a transitional stage; it is the important time when the child passes from pure vocal play to the very first utterance of linguistically interpretable sounds. S/he must restrict his/her large vocal possibilities to some linguistic and social constraints. This period has been the focus of a wide range of recent research which has an equally wide range of implications. On the one hand, there seems to be substantial evidence for universal development, which could reflect a maturational process independent of the language environment. On the other hand, there is equally compelling evidence for early language-specific influences on babbling. Both kinds of processes could easily have been predicted. However, there are somewhat mixed outcomes, and prosody has been overlooked. Therefore, in the present investigation, I shall focus on only one prosodic parameter, rhythm. But it should not be forgotten that time is only one of several possible components of rhythm, notably the prosodic feature of Fo is intimately associated with timing.

1. SYLLABLE-TIMED LANGUAGES.

1.1. French

1.1.1. French is generally said to have a "syllable-timed" tendency, because its syllables, mainly open, are more or less equal in duration. I prefer to describe it as being "trailer timed" [3] because each group or sentence ends with an accent whose main physical parameter is duration [4]. The final syllable (FS) is twice as long as the internal ones. Stress occurs on a whole string of words, and not on individual words. As its localization is imposed on right boundaries, the function of this final accent is clearly demarcative, indicating the end of a clause. The location of stress is thus completely predictable.

1.1.2. Subjects and Data Collection.

My population was composed of 12 babies who could hear well and had no birth problem, monitored from the age of 9; to 24;. Recordings were carried on each week during the critical period between 8 to 12; and once a month after on. All vocal productions except cries, emotive and vegetative sounds were included in the sample. The utterances were divided into strings defined on the basis of sequences separated by a pause of 400 ms or more.

For the pivotal stage, I used 16 hours of recordings, out of which 160 mn. have been analyzed instrumentally. It is the equivalent of the total amount of speech of a 2-year-old child in a 12 hour day. For the period 12-24; I selected 800 utterances for four babies followed longitudinally, and at least 40 utterances from the other children. A three level analysis was undertaken: 1) An auditory analysis made by 11 native listeners without any knowledge of the situation in which the child was while babbling, 2) An acoustic analysis in order to measure Fo, duration and location of stress, 3) A linguistic analysis to discover if intonation and rhythmic patterns had linguistic functions in the child's early productions.

1.1.3. The Syllabic Organization is totally different when the child is alone uttering non communicative vocalizations or when s/he is interacting with adults and emitting PL. At the ages of 9; and 10; non communicative utterances were mainly (71%) made up of vowel-like sounds. PL on the contrary, was mainly made up of canonical CV structures which can be reduplicated, or variegated. Structures with 2 or 3 syllables each represent 28% of the whole. Longer multi-syllabic utterances represent 29% of the whole. From now on, we shall examine only the temporal organization of these CV syllables of PL produced in interaction, so as to compare the data of the languages under investigation.

1.1.4. Temporal Organization.

The PL's CV structures have a short syllabic duration (M=250ms, s.d.105). It exceeds nevertheless by 30% the syllabic duration of adult speech. The dispersion is low, with a bell curve distribution.

One noticeable thing is the evolution of this temporal organization. First, at 9; the syllables are all nearly equal. A clear isosyllabicity exists at the beginning of the PL. After that, the duration of the syllables depends on their position in the utterance. The non-final syllables (NFS) gradually become shorter. The linear regression curve between age and duration of NFS becomes relevant from 10; on. FS have an unstable duration for quite a long time. However, as the NFS become shorter, the ratio FS/NFS becomes higher than 1.30, which gives the perception of final lengthening (FL).Later on, towards 16; the FS become much longer and finally are twice as long as the NFS. Of course, if one considers the details, the evolution can sometimes be more complex, and it even presents apparent regressions. In my data, variability in duration and the regression phenomenon appear especially in the FS. Contrary to the NFS, there is no significant correlation between age/FS, because the duration of SF is unstable through the different ages. Is this duration variability a problem of neuromotor maturation? If it were, the same thing should happen to NFS. I suggest that the greater variability in FS could be explained by the fact that the child tries to reach his/her target, which is a quite precise lengthening of the FS; as in learning to play tennis, sometimes the shots are too long, sometimes too short, which creates variability in duration. So, variability and regressions are only apparent on the surface level; they reflect a new organization at a deeper level. This is a major sign of the fact that final lengthening is being acquired, and has to be acquired. It is not a passive process; as in every acquisition strategy, there are errors and successes before one can attain the right target.

As a conclusion for French, I can say that my study has shown that the typical syllable structure and the typically trailer-timed rhythm of the French language, with its "point d'orgue" at the end of the utterance, are acquired in the middle of the second year. My results are consistent with others [1,5].

1.2. Hungarian

1.2.1. It is another syllable-timed language which can be compared to French because it has mainly open syllables and stress has a demarcative function: its location is predictable, it falls on the initial syllable and indicates the beginning an utterance or a word. The FS of an utterance is phonetically lengthened, but this lengthening does not have the same linguistic value as in French. However, there are two big differences in that open and closes syllables are nearly 50% each (CV 47, CVC 53%) in Hungarian and word order is constrained in French and free in Hungarian. In Hungarian, word order and thus stress location depend mainly on the speaker's intention, on topic/comment organization, and on given/new information. As a result, there can be more than one primary stress within a long utterance. It should also be added that in Hungarian the duration of Vs and of Cs is phonological, long Vs or long Cs being contrastive with short ones.

1.2.2. Child language data

The population was 1 female child followed from 9 to 36; and some children studied cross-sectionally by Kassai [6]. From her data we re-interpreted together (1995) for this paper, the following conclusions can be driven. The Hungarian child's speech contains many more open syllables than the adult speech. Until now, no precise count has been made on that subject. Concerning temporal organization, at the beginning there is an initial isochrony for the vowels, the phonological long ones and the short ones having the same mean values. Only much later on does the child shorten the phonological short vowels, the long ones keeping their initial duration. It also appears that the child uses stress from 12; on. However, the rules of stress patterning are not properly used. To know what is going on, one has to make a difference between one-unit utterances and longer utterances: a) In one-unit utterances, instead of putting a stress on the initial syllable, the child may use two different strategies: the first strategy is to stress only one syllable, which is correct in Hungarian, but it can be any syllable of the utterance. In reduplicated units, either both syllables are equivalent in duration (30%), or the first syllable is longer (30%), or the last syllable is longer (40%). In words

which are repeated, final stressing is strong . The second strategy is to stress more than one syllable, with no real preference for location. Even within a same word, stress can fall on different syllables from one occurrence to another; however, two successive syllables are seldom stressed, except in emphasis. Finally, the hierarchy of the different stresses is still missing at 19; the child stressing very often the first and the last syllable. This seems to show either a neutral start or a slight tendency towards FL.

b) In utterances containing more than one lexical unit, which the child uses towards 26 months, there are always more accents than there should be. Generally, instead of stressing only the new information, which is the rule in Hungarian adult language, the child stresses both the given and the new information. However, by this age, the child has found the correct initial location of the stress, but he may add a secondary stress on the FS. This behavior again shows a tendency towards FL, even in this language which has initial stressing; however, there is not much consistency in this FL, and this could also be interpreted as a neutral start, the child trying different possibilities, as if he were testing the functional value of the different locations and cues. Later on, the number of stressed syllables diminishes, stress falling mainly on the first syllable, with again, sometimes, a secondary stress on the last syllable.

Two noticeable things should be added: on the one hand, Hungarian words which are never stressed in adult language, are not stressed in child language either; on the other hand, words that are always stressed are also stressed by the child. These are very simple patterns, with no variability, and thus, their acquisition is not a problem.

2. STRESS-TIMED LANGUAGES.

such as English and Portuguese are said to have a "stress-timed tendency", because their basic rhythm is mainly determined by the stressed syllables, which tend to exhibit more or less regular interstress intervals (Pike 1946). English is the typical example of a stress-timed language; it has almost all the factors necessary to give the auditory impression of stress-timing. Stress location is not predictable, due to the fact that stress is lexical; it can fall on different syllables within a word, and has a strong grammatical function. Contrary to French, words keep their stress pattern within an utterance, which thus has more than one stressed syllable. FL is present in these languages, as a phonetic cue. Another point in the stress-timed languages is that the stressed syllable concentrates several sorts of prosodic features, and the unstressed syllables are shortened to the point that they can completely disappear; both intonation turning points and emphasis are on the stressed syllables. Stressed syllables are very prominent.

2.1. For the **English** language, we shall leave the description to M. Vihman [7], and give only a conclusion focusing on the main differences between French and English. From the literature, it appears that the typical closed CVC syllables of English appear very early: the English speaking child has already 2% CVC syllables at 8; going up to 10% at 11; and 25% at 14;. So, by the age of 10; English and French speaking babies have begun to produce the syllabic structure typical of their mother tongue. Comparing the temporal organization of French and English [8] it should be emphasized that these two languages, diametrically opposed in rhythm in adult language, seem to have quite similar syllable-timed rhythm in babbling and early speech; they differ nevertheless in their syllabic organization and in the consistency of FL. By the age of 16; the French child has already acquired the typically trailer-timed rhythm of his/her mother-tongue while the English-speaking child has not yet acquired the rhythmical pattern of his target language by the age of 2 years.

The questions now are: when and how does the English child acquire mastery of his/her language? When and how does the reorganization from syllable- or trailer-timing to stress-timing occur? A related question is: in which order are the rules of stress-placing learned? According to different scholars, location of stress is well perceived and also reproduced in experimental situations. However, in spontaneous speech, the correct stress pattern, with its grammatical function, does not seem to be properly used until the child is 3 years old, and generally quite later. This is beyond the age period of our study. The only clear thing that can be said is that from about 2;6 years the child is able to use the presence versus absence of prominence, but neither the stress location rules nor the hierarchy of the different stress patterns are acquired.

2.2. Portuguese although said to be stress-timed, shows only 3 or 4 of the 6 factors responsible for the stress-timing, i.e., vowel quality alteration, vowel reduction, compression of unstressed syllables, a large percentage of closed syllables, and relative flexibility in stress placement. However, the location of stress is far more predictable than in English; it can fall on different syllables, but the penultimate is the more often stressed. It should be added that intensity plays an important role [9].

2.2.1. Data for child language.

The study is not yet completed. Presently, we have a short follow-up, for 4 babies only, at 9; and 12; in good health, in very close interaction with their mothers. However, data is being collected from 100 babies from 2 months to 6 years recorded at the Lisbon Public Hospital, under the direction of Pr. Gomes Pedro. The study is both acoustic and perceptual.

2.2.2. Results.

At 9; 25% of polysyllabic utterances are made from CV's while at 10; these rise to 42%; mean duration at both ages is around 340 ms (min.100, Max. 950 ms). At both ages, the ratio between FS and NFS is between 1.22 and 1.13; thus, there is no perceivable FL, but a quasi-isochrony, and a great deal of variability. As it is well known that intensity is an important cue in Portuguese stress, an auditory analysis was made. This cue could not be studied instrumentally for technical reasons, the child's distance from the microphone being far too variable. The listeners were non Portuguese master's students in phonetics. They detected a prominence on the last syllable, presumably due to intensity in 73.6% utterances at 9;. However, at 12 months, these results were reversed: a prominence was detected on 71.8% of penultimate syllables. Does this mean that the typical rhythm of Portuguese is already acquired at this early age? We doubt it. These results may be explained by the fact that the child is in very close interaction with his/her mother, sitting on her lap, often trying to imitate the mother's model.

3. SPANISH belongs to the group having 'in-between' timing: its stress-timing tendency is not clear, mainly because closed syllables are seldom, and stress falls on the penultimate in 60% of the cases.

As data was taken from the literature [10], we shall go immediately to the conclusions.

For the 4 children followed between 19 and 26; there seemed to be an over-

all lack of preference for any kind of stress pattern. From an auditory analysis, the author concluded that one of the children showed a final stress procedure, while another preferred penultimate stress, and the two others had final stress in spontaneous tokens, but penultimate stress on imitated tokens. These results support a neutral-start hypothesis.

DISCUSSION.

This present study, along with previous papers and some cross-linguistic results taken from the literature, show that things are far from simple. On the contrary, what clearly appears is that the acquisition of rhythm, and hence stress patterning, is not as easy as is often said, by virtue of the principle that rhythm is inherent to all human activity, and that the child is already able to hear the rhythm of his mother tongue in utero.

My data suggest that an initial syllabic isochrony, followed by a more or less clear and stable FL or stressing is common to languages which have completely different stress patterning. So, we return to the hypothesis of the existence of a very general iambic rhythmic constraint due to an internal neural clock, with a regular rhythm, controlling the production at its base. This temporal structure may be governed only by biological rules. Its internal organization and its limits may correspond to the child's motor abilities. Nevertheless, these abilities have to be learned. They are neither innate, nor physiologically constrained, contrary to a widely held belief among researchers.. But if this temporal structure with its FL is not innate, it is interesting to consider it, with Lindblom [11] as a natural phenomenon found also in dance, music, insect stridulations, bird singing, Every temporally structured phenomenon seems to have a final lengthening, associated with the notion of ending. So lengthening and ending are in fact a consequence of the emergence of structuring. This does not exist at the very beginning of PL, because the child has not yet pre-programmed the whole utterance with its FL. Once the beginning of language structuring has appeared, FL, a cue of this structuring, appears too. Hence its presence in many languages. Acquisition of FL, which is after all a small phonetic detail, shows three major outcomes. 1) From a communicative point of view, it is an indication of good acquisition of turn-taking, and from then on, proto dialogs function very well. 2) From a cognitive point of view, the mapping of syllabic duration into the system shows the onset of a new stage in cognitive development marked by the appearance of a relational structure between the whole and its parts. 3) From a linguistic point of view, valid only for French, it shows that the child has integrated not only the overall rhythmical system, with each syllable having its own relative duration according to its position, but also its demarcative value.

As the child matures, FL will be superseded by accentuo-temporal patterning constraints specific to each language. Many phonetic, phonological, lexical, syntactic, and prosodic constraints will then prevent the internal neural clock from working correctly.

The last problem is to predict the emergence stages of these stress constraints. From our different studies, we suggest the following prediction: the rhythm of languages which have a "Gestalt" with natural FL and predictable stress location will be acquired early (in the first half of the second year). We have demonstrated that this is the case for French and also for some of the most simple stress patterns of Hungarian.

For languages which have a simple, frequent, and not very variable "Gestalt", with a syllabic structure of mainly open CV syllables where prominence is nearly stable, located near the boundaries, rhythm will be acquired a little later but still early, generally towards the end of the second year. This seems to be the case for Quiche Mayan for instance, or Mohawk, or Brazilian Portuguese (Stoel Gammon 1976), which is not a strongly stress-timed as continental Portuguese, or Comanche, whose stress is on the initial syllable as in Hungarian (Casagrande 1948).

When languages have a dominant, nearly predictable stress pattern, with some exceptions, the child, after a phase of FL, hesitates a while, chooses a strategy without stress preference, and then follows the patterning of his/her mother tongue. This may happen in the first half of the third year. It seems to be the case of Spanish. Portuguese, with its greater number of closed syllables and its numerous vowel reductions, could be a little bit more difficult, but the importance of the intensity cue may be a counterpart.

In languages which have no dominant "Gestalt", for neither syllabic structure, with a relatively high percentage of closed syllables, nor for prominence which is located in variable places depending on grammatical and lexical factors, the child, unable to find invariability and stability in the model, has more difficulty. Generally s/he begins to acquire part of the correct stress pattern only after two and a half years, when s/he combines 2 or 3 words. But the child makes many stress errors. English is a typical language with a long delay in acquiring the correct stress pattern; German also, to a lesser extent.

Now, from what we know about the rhythmical structure of different languages, our aim is to predict in which category each language belongs, and then to verify our predictions with perceptual and instrumental analyses. The final goal is to predict the errors a child will make in acquiring the stress pattern of his mother tongue.

[1] Allen, G., Hawkins, S. (1980), Phonological rhythm: definition and development. Yeni-Komshian et al. (ed), *Child Phonology*, vol.I, 227-256.
[2a] Konopczynski, G. (1990), *Le langage émergent: caractéristiques rythmiques*, Hamburg: Buske Verlag, 363p.
[2b] Konopczynski, G. (1993), Le bébé de deux ans a-t-il déjà acquis la structuration rythmique de sa langue maternelle? Besançon, *Bulag* 19, 73-85.
[3] Wenk, B. & Wioland, F.(1982), Is French really syllable-timed? *J. Phonetics*, 10, 193-216.
[4] Delattre, P. (1965), *Comparing the phonetic features of English, German, Spanish and French*, Heidelberg: J. Gross
[5]Levitt A. et al. (1991), From babbling towards the sounds systems of English and French: a longitudinal two-case study, *Haskins Laborat. SR*-107/108, 41-62.
[6] Kassai, I. (1991), The emergence of intonation and stress in Hungarian: a case study, *XII Int. Cong. Phon. Sc.*, 328-332.
[7] Vihman, M., *Phonological development: The origins of language in the child*, to appear 1995, Blackwell.
[8] Konopczynski, G. (1993), The phonological rhythm of emergent language: a comparison between French and English babbling. *Kansas Univ. Work. Pap. in Linguistics*, 18, 1-30.
[9] Delgado Martins, R. (1982), *Aspects de l'accent en Portugais*, Hamburg: Buske V.
[10] Hochberg, J. (1988), First steps in the acquisition of Spanish stress, *J. of Child Language*, 15/3, 273-292.
[11] Lindblom, B. (1978), Final lengthening in speech and music. Lund: Garding et al.(eds): *Nordic prosody*, 85-101.

ACQUISITION OF VOWEL DURATION: A COMPARISON OF SWEDISH AND ENGLISH

Carol Stoel-Gammon, Eugene H. Buder, and Margaret M. Kehoe
Speech and Hearing Sciences, University of Washington, Seattle, Washington, USA

ABSTRACT

This study compares durations of the high front unrounded vowels produced by 30-month-old subjects from two language communities: Swedish and American English. The findings indicate that intrinsic and extrinsic factors have different influences on vowel length in the two languages. In productions of children acquiring English, the extrinsic factor of voicing of the final consonant manifests strong effects on the vowel duration, but the intrinsic effects of the tense-lax distinction are absent. For Swedish children, the intrinsic effects are very strong, but there is no effect of consonant voicing.

INTRODUCTION

In acquiring the phonology of their mother tongue, children must learn not only the articulatory gestures needed for the correct production of the consonants and vowels of their language, but also the rhythm and timing associated with sounds, words and phrase.. Acquiring adult-like durational patterns is difficult in part these patterns vary considerably across speaker and context. The length of speech sounds, particularly vowels, is influenced by a number of intrinsic and extrinsic factors, including vowel quality; rate of speech; word and phrasal stress; position within the phrase; and consonantal context. Consequently, the child's is faced with more than simply learning a fixed duration for each segment or word.

In conjunction with researchers from Stockholm University, we have collected and are analyzing a large set of data from infants and toddlers aged 6-30 months. The goal of this joint research project is to examine the development of language-specific speech patterns in the two languages at both the segmental and suprasegmental levels. (See for example the comparison of acquisition of alveolar vs. dental /t/s in 30-month-old subjects from Seattle, USA and Stockholm, Sweden [1].) The investigation reported here focuses on aspects of the acquisition of vowel duration in the two languages.

Vowel Systems

Comparisons between the vowel systems of Swedish and American English are particularly interesting for crosslinguistic studies. The Swedish vowel system is typologically large, with up to 18 distinct vowels in contrast to the American English system which contains 12 vowel phonemes (excluding diphthongs). The Swedish system is typically described as being composed of nine short-long pairs, although quality differences in each pair do exist. [2,3]. The phonemic system of American English, by comparison, is generally described as having vowels that are distinguished primarily by differences in quality.

Previous investigations of vowel duration in the two languages have highlighted the presence of both intrinsic and extrinsic differences in duration. **Intrinsic** differences are readily apparent in tense-lax (or long-short) vowel pairs. For both languages, the vowel referred to as "tense" is longer than its "lax" counterpart in contexts in which other phonetic features are held constant. Thus, for example, in English the lax vowel /I/ of the word "bit" is shorter than the tense vowel /i/ of "beat". (Note: The IPA symbol [I] for the lax vowel used here corresponds to the Swedish Technical Alphabet short vowel i2.).

Extrinsic differences in vowel duration are conditioned primarily by phonetic features of the consonant which follows the vowel; if all other parameters are held constant, vowels tend to be shorter when they precede voiceless consonants than in other environments. Thus, in English, the /i/ of "beat" is shorter than the /i/ of "bead"). The finding that vowels are shorter before voiceless obstruents than before voiced obstruents and sonorants has been documented in a wide range of languages, leading some researchers to suggest that this particular durational difference is universal and physiologically determined (e.g., [3, 5].

However, Keating [6] points out that in a small number of languages, including Polish, Czech and Saudi Arabic, there is no interaction between vowel length and consonant voicing. Consequently, she argues that context-sensitive patterns of vowel duration represent rule-governed behaviors that are language specific and must be learned.

In the case at hand, American English and Swedish are among the large group of languages that follow the general pattern of longer vowels before voiced consonants. At the same time, as shown below, the two languages differ substantially in the magnitude of the effect on vowel duration.

Adult Data: Intrinsic Patterns

Detailed work by Elert [3] has shown that the intrinsic durational difference between lax and tense vowels in adult Swedish are substantial. The ratio of lax-to-tense vowels, averaged across all pairs, is .65, meaning that the lax vowels are about 2/3 as long as their tense counterparts. For American English, the durational differences are somewhat less marked with a lax-to-tense ratio around .71 according to House [4]. As noted above, in most phonological analyses of American English, intrinsic durational differences are considered to be secondary cues for distinguishing tense-lax pairs such as /i/ vs. /I/ or /u/ vs. /U/; the primary distinction stems from formant (i.e., quality) differences in such pairs.

Adult Data: Extrinsic Patterns

Data from existing studies of extrinsic vowel quantity effects indicate that voicing of the following consonant exerts a strong influence on vowel duration in American English. Vowels preceding voiced consonants are nearly twice as long as the same vowels before voiceless consonants. House [4] (see also Hoard [7]) reports that, on average, the ratio of vowels preceding voiceless compared with voiced consonants in monosyllabic words is .51.

In Swedish, the extrinsic effects of consonant voicing are much smaller, as might be expected in a language with a phonemic vowel length contrast. (If vowel durations were to shift substantially as a result of voicing of the following consonant, intrinsically long vowels might be perceived as their short/lax counterparts and vice-versa.) For adult Swedish, Elert [3] reports an average voiceless-to-voiced ratio of .97, indicating little influence of the voicing status of the following consonant.

If we compare the two languages, then, it is clear that intrinsic differences are strong in adult Swedish, and are present, but less marked in adult American English. Extrinsic differences in vowel length, by comparison, are very strong in American English, but are minimal in Swedish.

Child Data

To date, research on the acquisition of vowel length by children acquiring American English has provided inconsistent findings. Naeser [8] reported that the intrinsic and extrinsic vowel length ratios of her 22-month-old subjects were similar to those of the adults in her study. The children's intrinsic ratio was .74 (compared with an adult ratio of .71); the children's extrinsic ratio was .50 (cf. an adult ratio of .59). A year later, however, the children's ratios were less adult-like and differed minimally from one another:: The average intrinsic ratio was .68 and the average extrinsic ratio was .62. (The ratios given here were calculated on the basis of raw data presented in Naeser's report.)

In a subsequent study of vowel duration patterns in American children, aged 26 months, Greenlee [9] found that intrinsic and extrinsic duration patterns were approximately the same: .68 for the intrinsic ratio and .66 for the extrinsic ratio.

In sum, the data on acquisition of vowel duration by American children are unclear. At 22 months, the children in Naeser's study exhibited extrinsic differences that were more marked than

intrinsic differences, a pattern that conforms to adult speech. The findings for older children, however, indicate equal levels of shortening in both contexts, unlike adults.

We know of no data on the acquisition of vowel duration in children learning Swedish. Given the marked differences in the intrinsic and extrinsic durational patterns of Swedish and American English vowels, we might expect that adult-based patterns would emerge early in child speech. On the other hand, it is possible that all children will follow a similar developmental course, with intrinsic and extrinsic differences acquired in a fixed order.

Purpose

The purpose of this study was to examine the acquisition of intrinsic and extrinsic vowel length patterns in Swedish and American English by comparing productions of the vowels /i/ and /I/ in the speech of young children acquiring these two languages. The vowel pair /i/ - /I/ was selected for a number of reasons:

(1) In Swedish and English, these vowels form a tense-lax (or long-short) pair, making them a comparable match across the two languages;

(2) These vowels occur frequently in words that are part of young children's vocabulary and thus spontaneous productions could be elicited;

(3) Previous research suggests that these vowels are acquired relatively early in child speech [8,10].

Predictions

Given the findings of previous investigations of adult speech, we formed a set of predictions regarding the develop^mental patterns of intrinsic and extrinsic vowel duration in the speech of Swedish and American children:

(1) The vowels of *Swedish* children would be influenced more by intrinsic than extrinsic lengthening, as in the adult model.

(2) Extrinsic vowel length differences would be greater in the speech of *Swedish* children than in the models, because the children would have to learn to overcome the apparently universal pattern of lengthened vowels preceding voiced consonants

(3) Both intrinsic and extrinsic duration differences would be present in the productions of *American* subjects.

(4) Extrinsic differences would be greater than intrinsic differences for the *American* children, as they are in the adult model.

METHOD

Data Collection

Subjects included 18 children aged 30 months: nine acquiring Swedish in Stockholm; nine acquiring American English in Seattle. At both sites, the subjects participated in semi-structured tasks during which words with /i/ and /I/ followed by voiced and voiceless consonants were elicited. Care was taken to elicit monosyllabic forms with obstruent onsets and offsets to aid subsequent durational analyses. In addition, an effort was made to elicit phonetically similar words (including some nonsense forms) in the two languages.

All speech samples were recorded in sound-treated rooms using Lavalier microphones placed in a soft vest worn by the subjects. The microphone was linked to an FM wireless system. At both sites, the speech signals were recorded on Panasonic VHS videocassette recorders using High-Definition audio tracks.

Database

The data base for the present study consisted of productions of isolated or phrase-final words of the shape CVC in which the vowel was /i/ or /I/ and the inital and final consonants were obstruents. Because Swedish has few monosyllables in which lax /I/ is followed by a voiced consonant, few exemplars of this form were obtained; consequently this category was not included in the durational analyses for either language.

The children's productions were divided into three groups based on durational patterns:

Group 1 consisted of words with lax /I/ followed by a voiceless consonant;

Group 2 consisted of words with /i/ followed by a voiceless obstruent;

Group 3 consisted of words with /i/ followed by a voiced consonant.

Analyses of *intrinsic* durations are based on comparisons of Group 1 vowels with Group 2 vowels. Analyses of *extrinsic* duration involve comparisons between the vowels in Groups 2 and 3.

Acoustic Measures

All words were digitized and vowel durations were measured using the spectrographic display produced by the Computerized Speech Lab (Model 3400, Version 4.0 Kay Elemetrics). After eliminating tokens that were not acoustically analyzable because of poor voice quality or noise overlay, the vowel durations of 206 monosyllables were measured using the following criteria:

(1) Vowel onset was indicated by released vowel energy showing clear periodicity and energy in the first three formants.

(2) Vowel offset was indicated by the evidence of oral closure (i.e., a sudden reduction in waveform envelope and a loss of clear formant energy).

The first three formants of all vowel tokens were also measured using procedures developed by Buder and Stoel-Gammon [11].

RESULTS

Duration Measures

Average durations for the American and Swedish subjects for the vowels of interest are as follows.

Group 1 vowels: The mean duration of /I/ followed by a voiceless consonant was 202 ms for the American subjects and 161 ms for the Swedes.

Group 2 vowels: The mean duration of /i/ followed by a voiceless consonant was 191 ms for the American subjects and 326 ms for the Swedes.

Group 3 vowels: The mean duration of /i/ followed by a voiced consonant was 329 ms for the American subjects and 295 for the Swedes.

Intrinsic and Extrinsic Ratios

Figure 1 (next page) presents individual and group data on intrinsic and extrinsic vowel duration ratios. The top half of the figure shows the findings for the American subjects; the bottom half provides the same data for the Swedish subejcts.

Sections (a) present the individual ratios for extrinsic (Ext) and intrinsic (Int) durations for the nine subjects in each group;

Sections (b) present averaged group data for the individual ratios with error bars showing the standard error of the mean;

Sections (c) show formant measures (in mels) for the two vowels in question. /i/ is indicated by triangles and /I/ by circles.

The *intrinsic* ratios presented in Figure 1 are based on comparisons of /I/ and /i/ before a voiceless obstruent. The average intrinsic ratio for the American children was 1.06 [see Figure 1, Section (b), top half]. The average intrinsic ratio for the Swedish children was .49 [see Figure 1, Section (b), bottom half].

Average *extrinsic* ratios (based on subjects' averages of /i/ before voiceless and voiced obstruents) for the two groups were .58 for the American children and 1.10 for the Swedes [see Figure 1, Sections (b)].

Statistical Measures

Statistical analyses using the Wilcoxon rank sum test for dependent groups indicated that the extrinsic ratio in American English was significantly smaller than the intrinsic ratio ($p < .05$) [see Figure 1, Section (b), top half]. In Swedish, the intrinsic ratio was significantly smaller than the extrinsic ratio ($p < .05$) [see Figure 1, Section (b), bottom half].

Examination of the individual data in Figure 1, Sections (a) reveals that among the American subjects, two children weakly violated the group pattern by exhibiting intrinsic and extrinsic ratios that were essentially equivalent. Among the Swedish children, one subject strongly violated the pattern used by others in the group.

Quantity vs. Quality

The finding that there was little difference in the durational values for /I/ and /i/ in the productions of American children (with a lax-to-tense ratio of 1.06) raises the possibility that these vowels were indistinguishable in their speech. Acoustic analyses of formant structure indicated, however, that the children's /I/s and /i/s were clearly

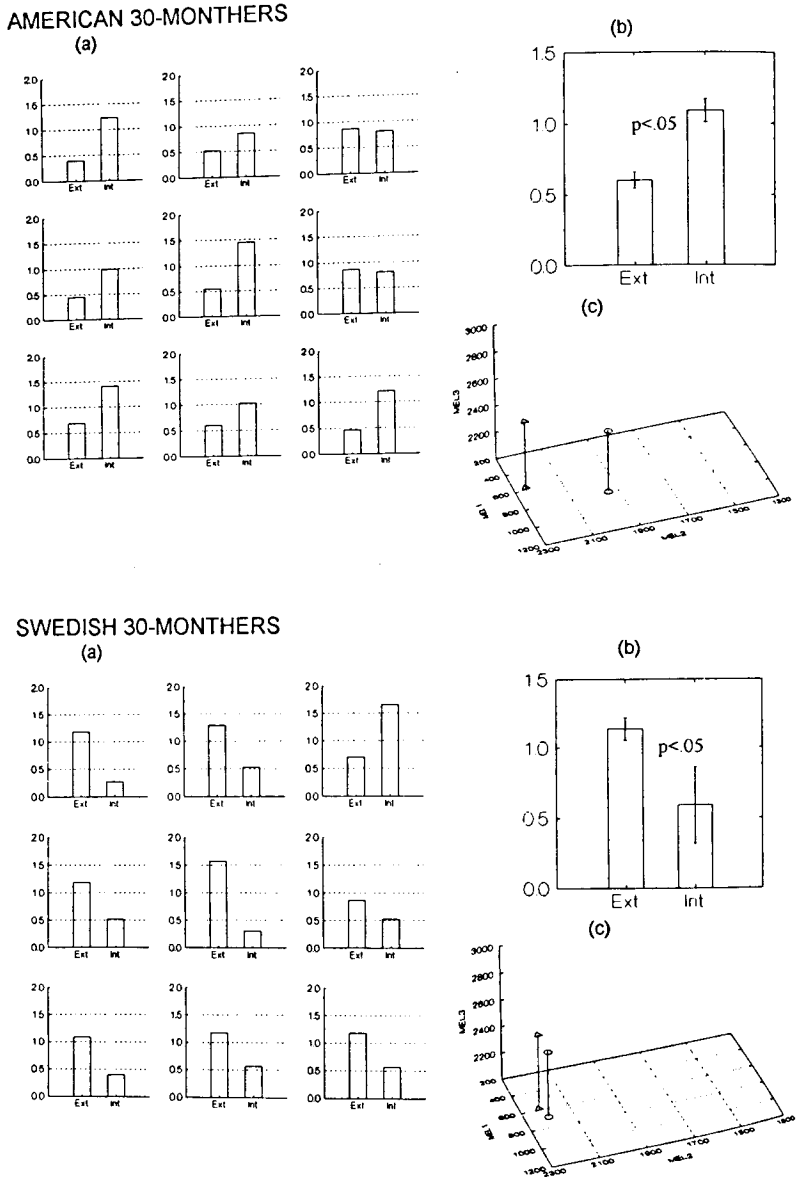


Figure 1. Comparison of quantity and quality measures of /i/ and /I/ from CVC words produced by 9 American (top panel) subjects , aged 30 months, and 9 Swedish subjects (bottom panel) aged 30 months. Sections (a) show individual data; sections (b) and (c) show group averages. See text for a full explanation.

different in terms of quality. As shown in panel (c) in the upper half of Figure 1, the /i/ tokens produced by the American subjects were characterized by a lower F1 and a higher F2 than /I/, as is the case in adult speech.

In contrast, the /I/ and /i/ productions of the Swedish children exhibited little difference in quality, as shown in the lower half of Figure 1, Section (c). Rather, the vowels were distinguished by differences in quantity -- on average, /i/ was more than twice as long as /I/.

DISCUSSION

Comparisons with Adult Data

If we compare the children's ratios with those of adults, we see that the patterns are roughly similar, though certainly not identical. For adult Swedish, Elert [3] reported an ratio of .68 for /I/ and /i:; the children's ratio from this study was .49. The extrinsic ratio for /i/ before voiceless and voiced consonants in adult speech was .99 according to Elert, compared with 1.10 for the Swedish children.

For American English, House [4] reports an adult intrinsic ratio for /I/ vs /i/ in monosyllables of .74, compared with the children's ratio of 1.06. The adult extrinsic ratio for /i/ before voiceless and voiced consonants was .49 for adults and .58 for children in this study.

Predictions

Four predictions regarding vowel length patterns in the speech of young Swedish and American children were made at the outset of the study. Some were supported; other were not. The predictions are repeated below and compared with the findings.

(1) *The vowels of Swedish children would be influenced more by intrinsic than extrinsic lengthening, as in the adult model.*

This prediction was strongly supported. The children's productions exhibited a very strong influence of intrinsic shortening; in fact, the influence was much stronger in the children's productions than in the adults from Elert's study.

(2) *Extrinsic vowel length differences would be greater in the speech of Swedish children than in the adult models, because the children would have to learn to overcome the apparently universal pattern of lengthened vowels preceding voiced consonants.*

This prediction was not supported. On average, the Swedish children produced longer vowels before voiceless consonants than before voiced ones, in spite of the fact that most languages of the world conform to the predicted pattern. This finding raises questions about the physiological underpinnings of vowel lengthening preceding voiced consonants.

(3) *Both intrinsic and extrinsic duration differences would be present, in the productions of American subjects.*

This prediction was not supported. In the speech of the American children, the lax vowel /I/ was, on average, slightly longer than the tense vowel /i/, contrary to the adult pattern. As noted above, the primary distinction between /I/ and /i/ was in quality rather than quantity.

(4) *Extrinsic differences would be greater than intrinsic differences for the American children, as they are in the adult model.*

This prediction was supported.

Future Research

The findings from this study are intriguing and suggest several avenues of future research:

(1) Studies of younger and older children are needed to trace the development of language-specific durational patterns from emergence to mastery.

(2) Investigations of other vowels are needed to determine if the findings for the pair /I/ - /i/ hold true across the vowel system. Of particular interest in this regard would be an examination of other Swedish pairs that differ to greater extent in quality. It may be that durational differences are acquired later in such pairs since phonmic contrast is marked by quality az well as quantity.

(3) Information on individual patterns of development would be useful in comparing the roles of quality and quantity in the acquisition of the vowel system. For example, in this study, the durations of one Swedish child differed dramatically from those of her peers. It is possible that she was using vowel

quality rather than quantity to create vowel contrasts.

(4) Data on the acquisition of vowel durations in disyllabic words are needed for a full understanding of the influence of phonetic context and word length on intrinsic and extrinsic patterns.

(5) Finally,investigations of adult-child interactions are needed to determine the durational patterns of vowels in child-directed speech.

ACKNOWLEDGMENTS

This research was supported in part by a grants from NIH-NIDCD, the Royalty Research Fund and the Virginia Merrill Bloedel Hearing Research Center at the University of Washington. Data collection was carried out in collaboration with Karen Williams and Olle Engstrand of the Institute of Linguistics, University of Stockholm.

REFERENCES

[1] Stoel-Gammon, C., Williams, K., & Buder, E.H. (1994) Crosslanguage differences in phonological acquisition: Swedish and American /l/. *Phonetica,* vol. 51, pp. 146-158.

[2] Fant, G. (1973). *Speech sounds and features.* Cambridge, MA: MIT Press.

[3] Elert, C-C. (1964) *Phonological studies of quantity in Swedish.* Uppsala: Almqvisst & Wiksells.

[4] House, A.S. (1961). On vowel duration in English. *The Journal of the Acoustical Society of America,* vol. 33, pp. 1174-1178.

[5] Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica,* vol. 22, pp. 129-159.

[6] Keating, P.A. (1984). Phonetic and phonological representation of stop consonant voicing. In V. Fromkin (Ed.), *Phonetic linguistics.* New York: Academic Press.

[7] Hoard, J. E. (1966). *Contrastive analysis of English and Swedish phonology.* Unpublished master's thesis, University of Washington, Seattle, WA.

[8] Naeser, M. A. (1970). *The American child's acquisition of differential vowel duration.* (Tech. Rep. No. 144). Madison, WI: University of Wisconsin, Center for Cognitive Learning.

[9] Greenlee, M. (1977). *Another look at the acquisition of differential vowel duration in American English.* Unpublished doctoral dissertation, University of California, Berkeley.

[10] Otomo, K. & Stoel-Gammon, C. (1992). The acquisition of unrounded vowels in English. *Journal of Speech and Hearing Research,* vol. 35, pp. 604-616.

[11] Buder, E., & Stoel-Gammon, C. (1993). Obtaining valid and reliable acoustic measures of children's vowel productions. Paper presented at the American Speech-Language-Hearing Association, Anaheim, CA, November.

PARTICLES AND PREPOSITIONS IN SCANDINAVIAN CHILD LANGUAGE DEVELOPMENT: EFFECTS OF PROSODIC SPOTLIGHT?

Sven Strömqvist
Dept. of Linguistics,
University of Göteborg,
Sweden

Ann Peters
Dept. of Linguistics,
University of Hawai'i,
U.S.A.

Hrafnhildur Ragnarsdóttir
University College of
Education, Reykjavík,
Iceland

ABSTRACT

The syntactic and prosodic properties of particles and prepositions vary within the group of the Scandinavian languages in ways that offer a testing ground for the Prosodic Spotlight Hypothesis. This hypothesis predicts that elements that are made perceptually prominent by virtue of prosodic traits (stress, pitch, duration, rythmic patterns etc) will be focussed on earlier in language development than elements not so spotlighted. The paper discusses evidence from Danish, Icelandic and Swedish child language development.

INTRODUCTION

The world's languages all employ pitch, duration, and some kind of rhythm in their individual prosodic systems. These tonal and temporal characteristics not only give shape to utterance contours, and perform discourse–related functions, they also interact with grammar in ways that may have interesting consequences for both processing by adult speakers and learning by children. The specific kinds of interrelations between prosodic features on the one hand and aspects of lexical and grammatical structure on the other, vary a good deal across languages, however.

In an earlier study [11] two of us (Peters and Strömqvist) explored the idea that the prosodic patterning characteristic of a particular language can indeed serve to draw the attention of language learners to the presence of certain elements of the linguistic system (see also [10]). Awareness of the *presence* of such a form may then help focus the learner's attention on its other attributes, including exactly what it sounds like and what functional role(s) it has. More specifically, we proposed the following "Spotlight Hypothesis":

> Perceptually salient prosodic patterns, including pitch contours, rhythm, and increased duration, may serve as "spotlights" on any phonological forms that are regularly associated with these patterns; if such forms happen to be grammatical morphemes, learners will focus on them earlier than on morphemes not so spotlighted.

The Spotlight Hypothesis thus concerns children's *perception* of salient prosody that fortuitously coincides with grammatical morphemes, with evidence to be drawn from what children *produce* and from the parental input they receive. The Spotlight Hypothesis represents an attempt to bridge the gap between studies of infant perception (focusing on the first year of life) on the one hand and studies of early grammatical development (typically, from 18 months and onwards) on the other.

In our earlier study we explored in some detail the interaction between the Swedish grave word accent contour, i.e., the marked member of the Swedish tonal word accent distinction (see [2]; [3]; [4]), and the first inflectional morphemes in the early language development of a Swedish child between 15;19 and 30;20. In the adult target language, the distribution of the Swedish tonal word accents (acuteness versus graveness) can, to a large extent, be predicted from morphological information, that is, the phonetic gestures interact with grammatical information.

During a first phase, the child (re)produced inflectional morphology predominantly in utterance–final position and he overgeneralized the grave accent, especially the post–stress rise/high pitch, to most word forms with a post–stress syllable, including those forms where the post–stress syllable encoded an inflectional morpheme. In a second phase, starting around the same time as the child had productively acquired his first small set of inflectional morphemes, he withdrew the grave accent from these forms (resulting in an undergeneralization). During this second period, he produced inflectional morphemes with increasing frequency in the less salient non–utterance–final positions. In a third phase, the child acquired a distributional pattern of graveness which approximated that of the adult target.

These observations, especially the findings from the first phase of the longitudinal case study, are in accordance with Engstrand et al. [5], who, on the basis of an experimental study of children's early production data, argue that Swedish children already begin to master the phonetic aspects of the Swedish grave accent, especially the high pitch on the post–stress syllable, around 17 months of age, that is, well before they start acquiring inflectional morphology. The grave contour, especially the perceptually salient high pitch/post–stress rise, thus represents a phonetic gesture which is established both in the child's perception and production during his pre–grammatical development. It is therefore available to serve as a spotlight on elements which can be useful in the extraction and construction of morphosyntactic patterns.

Faced with the task of learning a language with a fair amount of grammatical morphology located at the ends of words, the Swedish–learning child does well to attend to prosodic salience which spotlights what goes on in this position. Such a strategy has been described by Slobin [12], p 335, in his Operating Principle "pay attention to the ends of words". On the basis of the findings from one longitudinal case study [11] we concluded that Swedish is a particularly felicitous language for learners to apply this principle because of the presence of prosodic spotlighting (increased duration and high pitch) on final syllables which also happen to be segmentable grammatical morphemes. The increased duration is due to the cross–linguistically attested final lengthening effect (see [7]; [6]), whereas the high pitch is due to the particularly Swedish grave post–stress rise.

The present paper extends the testing of the Spotlight Hypothesis to the acquisition of particles and prepositions in Scandinavian languages. Particles and prepositions belong to a small set of phenomena where these languages, which are otherwise typologically very similar, differ in terms of syntactic distribution and prosodic prominence. "The natural linguistic laboratory" of Scandinavian languages has, as it were, set slightly different scenes for young language learners in the area of particles and prepositions. In order to explore the possible effects of these differences, data were were drawn from a current inter–Nordic project, "Language Development — a Scandinavian Perspective" (see [14]).[1] The contrastive developmental anal-

[1] More precisely, the analyses presented in this paper relate to two Danish, two Swedish and one Icelandic longitudinal case studies, all collected in everyday situations in the home. The Danish and Swedish material is accessible in CHILDES/CHAT format (see [9]; [8]). Users of Internet can access a large set of CHAT–files from the Danish and Swedish child language corpora through anonymous ftp to poppy.psy.cmu.edu, where they are stored in the tar files "Danish.tar" and "Swedish.tar" under the direc-

yses were confined to particles and prepositions with a spatial (as opposed to temporal or general grammatical) meaning, all in order to reduce the number of factors that might influence the structure of acquisition. And in all five children alike, the first handfull of particles and prepositions that emerged in development were chosen from the same narrow range of options — *'in', 'on', 'up', 'down', 'out', 'off'* — encoding the same or very similar spatial concepts. However, the children varied considerably in terms of timing of acquisition as well as in terms of how many items they had acquired at an early age. Moreover, this variation showed language specific effects. Our evidence comes from two separate substudies, the first concerned with the acquisition of verb particles in Danish and Swedish, the second with the acquisition of prepositions in Icelandic and Swedish.

PARTICLES: DANISH VERSUS SWEDISH

The minimal variation between the Mainland Scandinavian languages (Danish, Norwegian and Swedish) includes systematic differences in the syntax and prosody of the VERB + PARTICLE construction (see [13]). If we focus on transitive verb phrases where the object is a pronoun, we get the situation summarized in table 1. The phrase used to illustrate the variation is TAKE *(ta)* IT *(det)* OUT *(ud/ut)*.

properties of PRT	SWEDISH	DANISH	NORWEGIAN
	ta **út** *det*	*ta det* **úd**	**tà** *det* **ut**
Contiguity with V	+	–	–
Phrasal stress	+	+	–
Phrase final position	–	+	+

Table 1: Differences in the VERB + PARTICLE construction between the minimally different languages Swedish, Danish, and Norwegian

Now, if we focus only on the first parameter in the table, syntactic contiguity of the particle with its verb, we would predict (a) that Swedish children will have an easier time (than Danish or Norwegian children) of perceiving the close connection between particle and verb, since the two are delivered together in the input to the child. If, however, we focus on the second and third parameters, we predict (b) that the particle will be maximally easy to attend to in the Danish case, where it receives both stress and extra prosodic prominence by virtue of its phrase final position. And in cases where this phrase final position coincides with utterance final position, the Danish child can also profit from the final lengthening effect which gives him extra time to perceive the particle. In contrast, Norwegian children are expected to have the hardest learning task according to the first and second parameters in the table.[2]

For the purpose of empirical testing, the syntactic distribution of the first six grammatical morphemes encoding spatial relations (that is, *'in', 'on', 'up', 'down', 'out', 'off'*) in the two Danish case study materials ("Jens" and "Anne") were compared to the corresponding distributions in the two Swedish case studies ("Markus" and "Harry"). The distributional analyses were made 1) in terms of timing (age of appearance across available data points) and 2) in terms of whether the grammatical morphemes occurred as one-word utterances or as elements of multi-word utterances. The analyses focussed on the first 20 data points (transcripts) available from each child/case study material. The results are presented in table 2.

(Insert table 2 here)

tory "/noneng".

[2] We have not yet started to analyse Norwegian data, but this is a priority for our future research.

Danish, Jens, first 20 data points		
Period	1-word utterances with spat adv/prt	multi-word utterances with spat adv/prt
12;26	0	0
13;23-18;26	8	0
19;14-22;14	11	7
22;28-24;02	0	31

Danish, Anne, first 20 data points		
Period	1-word utterances with spat adv/prt	multi-word utterances with spat adv/prt
13;01-18;20	49	1
19;04-22;17	69	44
23;18-23;26	1	22

Swedish, Markus, first 20 data points		
Period	1-word utterances with spat adv/prt	multi-word utterances with spat adv/prt
15;19-20;05	0	0
21;07-22;25	0	18[a]
23;00-27;28	2	360

Swedish, Harry, first 20 data points		
Period	1-word utterances with spat adv/prt	multi-word utterances with spat adv/prt
18;20-23;18	0	0
24;16-32;27	0	181[b]

Table 2: The distribution of particles on one-word vs multi-word utterances in the early language development of two Danish and two Swedish children

[a] 12 (67%) of which occur immediately after a verb

[b] 104 (57%) of which occur immediately after a verb. 100% of the tokens of *in* and *upp* occur immediately after a verb.

child	language	period	data points	particles (tokens)	% utterances with particles	% prt in final pos.
Anne	Danish	13;01-18;20	11	539	11.4%	15.8%
Markus	Swedish	15;19-23;00	10	293	14.3%	7.1%

Table 3: Distribution of particles in the early input to Anne (Danish) and Markus (Swedish)

The analysis shows that particles first emerge as one–word utterances in the Danish children (13–18 months of age). And when the Danish children begin to produce them in multi–word utterances, they tend to combine them with words other than verbs. In contrast, particles almost never occur as one–word utterances for the two Swedish children, and they are initially combined just with verbs in a clear majority of cases.[3] These two results render support to our first prediction (a).

The analysis further shows that particles emerge much earlier in the development of the two Danish children (around 1 year of age) than in the development of the two Swedish children (around 2 years of age). This finding renders support to our second prediction (b).[4]

Predictions (a) and (b) above rely on the assumption that the structural contrasts summarized in table 1 for VERB + PARTICLE constructions are reflected in the input heard by the Danish and Swedish children in our study. We have just started to test this assumption and results are available for Anne (Danish) and Markus (Swedish). Table 3 shows the distribution of particles in the input to Anne and Markus in the early phase(s) of acquisition evidenced in the two case studies. The table summarizes the number of data points analysed for each child, the number of particles found in the input utterances, the percentage of input utterances that contain a particle, and the percentage of particles that occur precisely in utterance final position.

(Insert table 3 here)

The table shows that

- for the Danish and Swedish case studies the proportions of input utterances that contain particles are very similar, and that
- there is a higher proportion of particles in utterance final position in the Danish input

These observations thus provide support for our assumption that this minimal but crucial syntactic difference between Danish and Swedish is already reflected in speech to children at an early stage of acquisition.

PREPOSITIONS: ICELANDIC VERSUS SWEDISH

In Danish and Swedish, just as in English, prepositions are unstressed and verb particles are stressed. The phonological forms occurring as particles are thus prosodically spotlighted and can, in effect, be expected to be attended to and internalized earlier by the child than forms occurring as prepositions. Further, some phonological forms can occur both as prepositions and as particles, e.g., *i* 'in' and *på* 'on'. And we find across the two Danish and the two Swedish case studies we have analysed so far, that, indeed, the first prepositions to emerge in the children's production are forms that also occur as particles in the specific input to these children. Let us assume that children first establish the phonological form of a particle/preposition on the basis of its occurrences in stressed (i.e., particle) position. They can then use this information to help them recognize these forms when they occur as unstressed phrase–internal prepositions. On this account we would expect that it is precisely these dual particle–prepositions that will be the first prepositions produced in early grammatical development.

We observed above that the first handfull of particles in Scandinavian child language development also includes the adverbs *in* 'in$_{dir}$', *upp* 'up', *ner* 'down', and *ut* 'out'. If we turn to Icelandic, the corresponding items are ambiguous between adverb and preposition and their status is determined by context. When they occur in a verb phrase like *hljóp út*[5] 'ran out' they are classified as adverbs according to Icelandic grammatical descriptions (see, e.g., [15]). In this type of construction both the verb and the adverbial element receive stress, that is V́ + PŔT.[6] And when they occur in unstressed position before NP, they count as prepositions, for example *hljóp út ganginn* 'ran out (of) the corridor', that is, V́ + PREP + N̄P. A consequence of this set of distributional properties is that the great majority of phonological forms that can serve as prepositions (unstressed) can also appear in stressed position, namely when they are used as adverbs. In Swedish (or Danish) the class of items of which the same distributional properties are true constitutes but a minority (basically, it is confined to a subset of compound prepositions, such as, e.g., *framåt* 'right-on-towards' and *igenom* 'in-through'). Again, on the assumption that children adopt the strategy of establishing the phonological forms in question on the basis of their occurrences in stressed (rather than unstressed) position, Icelandic children would be able to apply this strategy to a greater number of items than their Swedish or Danish peers.

To explore this hypothesis empirically, the set of prepositions which the longitudinal subject "Markus" produced by 24 months of age was compared with the corresponding data from the Icelandic subject "Ari". Again we restrict our analyses to items encoding spatial relations.

The difference found in the production data from Ari (Icelandic) and Markus (Swedish) at 24 months is striking: 12 different prepositions for Ari and only 3 for Markus.[7] There is no corresponding difference in the input data to the two children; in fact, the input data are identical in terms of type frequency: 23 different prepositions in the input to Ari and 23 to Markus. There is a striking difference between the Icelandic and the Swedish input, however, in terms of how the 23 items are distributed over positions with and without stress. Out of the 23 Icelandic items 12 were also used as adverbs. That means that the phonological forms of 12 of the 23 prepositions occurred with stress. The corresponding figures for Markus is just 3 of 23 (3 out of 23 prepositions were also used as particles). The three prepositions in question were the same three as Markus produced in his own speech by 24 months of age, i.e., *i* 'in', *på* 'on', and *till* 'to'.

CONCLUSIONS

We interpret the precocious emergence of particles in the two Danish children and the likewise precocious acquisition of prepositions in the Icelandic child as effects of prosodic spotlight, — although not exclusively of prosodic spotlight. In a controlled

[3] During a first phase, 67% of the particles used by Markus occur immediately after a verb. In Harry, the distribution of particles in contexts with verbs is more governed by the particular morphemes: 100% of the tokens of *in* 'IN$_{dir}$' and *upp* 'UP$_{dir}$' occur immediately after a verb, whereas only 15% of the tokens of *i* 'IN$_{loc/dir}$', occur with a verb.

[4] An additional factor which probably contributes to the precocious emergence of particles in the Danish children is recency: since the particle occurs in phrase final position in Danish, it is subjected to the so–called recency effect, which makes it easier to remember elements occurring at the end of linear structures (see, e.g., [16]).

[5] The diacritic signs in the Icelandic examples are there for orthographic reasons and are not related to stress.

[6] In Icelandic, the construction V + PŔT (i.e., with phrasal stress on the particle) is reserved for lexicalized (non–compositional) meanings, for example, *líta út* 'look'.

[7] Ari's 12 prepositions were *í* 'in', *inní* 'inside/into', *oní* 'down-inside/onto', *á* 'on', *niðrá* 'down–on', *uppá* 'up–on/onto', *af* 'of', *til* 'to', *frá* 'from', *hjá* 'by/with', *undir* 'under', and *yfir* 'over'. Markus's 3 prepositions were *i* 'in', *på* 'on', and *till* 'to'. There are no indications that this difference follows from an overall difference in grammatical level between the two children such that Ari would be more advanced than Markus. Across the data points from Ari at 24 months, Ari's MLU in terms of number of words per utterance is 2.33, that is, indicative of Brown's "stage 2" speech [1]. The corresponding figure for Markus is MLU 3.11.

experiment in which the task is to learn a fragment of an artificial language and the subjects of the experiment group would profit from the presence of prosodic spotlight, it would be possible more clearly to disentangle such a spotlight from other determining factors (such as input frequency etc). Real language development, however, takes place in multidimensional environments where a host of factors interact to determine the structure of acquisition. The purpose of our particular crosslinguistic approach, the within-group approach, is to study aspects of language development in naturalistic settings, keeping as many factors as possible under control, while varying the particular determining factor under scrutiny. On the basis of our intra-Scandinavian contrastive analyses, we conclude, then, that prosodic spotlight can interact with other determining factors (such as cognitive development, input frequency, etc) in ways that facilitate the acquisition of particles and prepositions to a degree that is clearly observable. Our observations indicate that this facilitation process can already be evident at the beginning of the child's second year of life.

Acknowledgments

This study was supported by Nordiska Samarbetsnämnden för Humanistisk Forskning (NOS-H). Thanks go to Helga Jonsdóttir, Åsa Nordqvist and Ulla Richtoff for valuable discussions and help with the coding work.

References

[1] R. Brown. *A First Language: The Early Stages.* Allen and Unwin, London, 1973.

[2] G. Bruce. *Swedish Word Accents in Sentence Perspective*, volume 12 of *Travaux de l'Institut de Linguistique de Lund.* CWK Gleerup, Lund, 1977.

[3] O. Engstrand. F_0 correlates of tonal word accents in spontaneous speech. *PERILUS*, 10:1–12, 1989.

[4] O. Engstrand. Phonetic features of the acute and grave word accents: data from spontaneous speech. *PERILUS*, 10:13–37, 1989.

[5] O. Engstrand, K. Williams, and S. Strömqvist. Acquisition of the Swedish tonal word accent contrast. Actes du XIIème Congres International des Sciences Phonétiques, Aix-Marseille: Publications de l'Université de Provence, pages 324–327, 1991.

[6] P. Hallé, B. de Boysson-Bardies, and M. Vihman. Beginnings of prosodic organization: Intonation and duration patterns of disyllables produced by Japanese and French infants. *Language and Speech*, 34:299–318, 1991.

[7] B. Lindblom. Final lengthening in speech and music. In E. Gårding, G. Bruce, and R. Bannert, editors, *Nordic Prosody*, pages 85–101. Department of Linguistics, University of Lund, 1978.

[8] B. MacWhinney. *The* CHILDES *Project — tools for analyzing talk.* Erlbaum, Hillsdale New Jersey, 1991.

[9] B. MacWhinney and C. Snow. The child language data exchange system: An update. *Journal of Child Language*, 17:457–472, 1990.

[10] A.M. Peters. Language typology, individual differences and the acquisition of grammatical morphemes. In D. I. Slobin, editor, *The Crosslinguistic Study of Language Acquisition Vol 4.* Lawrence Erlbaum, Hillsdale, New Jersey, 1995.

[11] A.M. Peters and S. Strömqvist. The role of prosody in the acquisition of grammatical morphemes. In J. L. Morgan and K. Demuth, editors, *From Signal to Syntax.* Lawrence Erlbaum, Hillsdale, New Jersey, 1995.

[12] D. I. Slobin. Developmental psycholinguistics. In W.O. Dingwall, editor, *A Survey of Linguistic Science*, pages 298–400. Univ. of Maryland Press, Univ. of Maryland, 1971.

[13] S. Strömqvist. Nordens språk som förstaspråk. In M. Axelsson and Å . Viberg, editors, *Nordens språk som andraspråk*, pages 115–132. Stockholms Universitet, Stockholm, 1992.

[14] S. Strömqvist, H. Ragnarsdóttir, O. Engstrand, Helga Jonsdóttir, E. Lanza, M. Leiwo, Åsa Nordqvist, Ann Peters, K. Plunkett, Ulla Richtoff, H. G. Simonsen, J. Toivainen, and K. Toivainen. The inter-Nordic study of language acquisition. Nordic Journal of Linguistics *in press*, 1995.

[15] H. Thráinsson. *Setningafræði.* Málvísindastofnun Háskóla Íslands, Reykjavík, 1990.

[16] A. Wingfield and D. L. Byrnes. *The Psychology of Human Memory.* Academic Press, New York, 1981.

THREE-DIMENSIONAL ULTRASOUND AND MAGNETIC RESONANCE IMAGING: A NEW DIMENSION IN PHONETIC RESEARCH.

A. K. Foldvik, Dept. of Linguistics, University of Trondheim, Norway
U. Kristiansen, Dept. of Acoustics, University of Trondheim, Norway
J. Kværness, MR-Centre, Medical section, Trondheim, Norway
A. Torp, Dept. of Computer Systems and Telematics, University of Trondheim, Norway
H. Torp, Dept. of Biomedical Engineering, University of Trondheim, Norway

ABSTRACT

This is an overview of the use of magnetic resonance imaging and ultrasound to produce three-dimensional models of the tongue and the vocal tract.

INTRODUCTION

In phonetic descriptions of sounds we are all used to midsagittal views of the vocal tract showing tongue shape and distance from the roof of the mouth. If a sagittal plane e.g. 15mm away from the mid-line for the same sound was chosen, we would get a very different view. Those midsagittal figures are convenient for pedagogical purposes but give no information about the shape and position of the tongue off the midline, e.g. bunching or flattening of the sides of the tongue, and may therefore not necessarily capture important differences between two articulations which mid-sagittally may look quite similar.

Many researchers with an interest in physiological phonetics, who so far have not been too impressed with acoustic vocal tract models, are working on supplying accurate three-dimensional models of the oral and nasal cavities using different techniques.

Even if x-ray techniques used for phonetic research have low radiation levels, techniques with no known health risks like ultrasound and magnetic resonance imaging (MRI) attract interest among researchers in many countries. We will give some examples of advances with these two techniques in phonetic research during recent years.

MRI.

Rokkaku, Hashimoto, Imaizumi, Niimi and Kiritani [1] were among the first to publish phonetic research results based on MRI. The MR images from these first years were fuzzy and were results of dauntingly long acquisition times where subjects would often try to hold an articulatory posture for more than one minute while lying on their backs in the narrow and extremely noisy MR tunnels of that time. The resulting fuzziness in the pictures was an aggregate result of inability to keep the articulators still, too thick picture cuts which might also be laid at acute angles of e.g. the tongue, and acquisition times which ideally should have been even longer to produce sharper images. With recent progress in MR-technology acquisition times of 1s or less for good quality images with a 5mm cut are unproblematic. This means that prolongable sounds like vowels, fricatives, laterals and nasals can be studied by means of MR.

In principle the picture plane, normally 5mm thick, can be placed at any angle, but to increase sharpness it is important that it is laid as close to 90° as possible on the surface of the structure to be imaged.

At first sight an x-ray and MR-picture look similar, but structures which contain little or no hydrogen like teeth and bones do not show up on a MR-picture, while cartilages and particularly soft tissues like the velum and tongue do. One small advantage in favour of MR is that dental fillings which tend to obscure the shape of the tongue in x-ray pictures do not show up on an MR-image. Yang and Kasuya [2] ingeniously solved the problem with no-showing dentition: They put dental impressions of the subjects in water, and took very accurate MR-images of the impressions. Coating mediums of the teeth have been tried with no great success so far.

Recently both mid-sagittal MR films of articulatory movements as well as dynamic three-dimensional vocal tract models by means of MRI have been made [3, 4]. See fig. 1, 2 and 3.

In X-ray technique enormous progress has been made. When we compare the improvements in x-ray techniques during nine decades from e.g. the misty pictures that Grunmach [5] published in 1907 to the razor-sharp xeroradiographic photographs of today [6] with progress in MRI during less than one decade, it does not seem too risky to predict that we will in the not too distant future see improved quality MR images of reduced slice thickness based on shorter acquisition times than today. We also envisage 3D films of articulatory movements using improved computer algorithms for air-tissue boundary detection.

ULTRASOUND

Compared to using MRI, ultrasound equipment is from our experience cheaper and easier to get access to, but there are limits to what you can do with it. Therefore anybody who has tried to record tongue movement and tongue shape for speech sounds by means of ultrasound must be impressed with Stone and Lundberg [7] who have presented three-dimensional reconstructions of the tongue surface based on speech sounds which were sustained for 10s.

One of the limitations to ultrasound equipment is that it does not register the tip of the tongue if the apex is so far forward in the mouth that the ultrasound transducer waves are unable to reach it, but instead register the sublingual air/tissue borderline. Also edge/surface detection is not always unproblematic. Sounds produced with contact between the tongue and the roof of the mouth pose another problem since tongue edge detection then becomes difficult. Therefore we have so far only succeeded

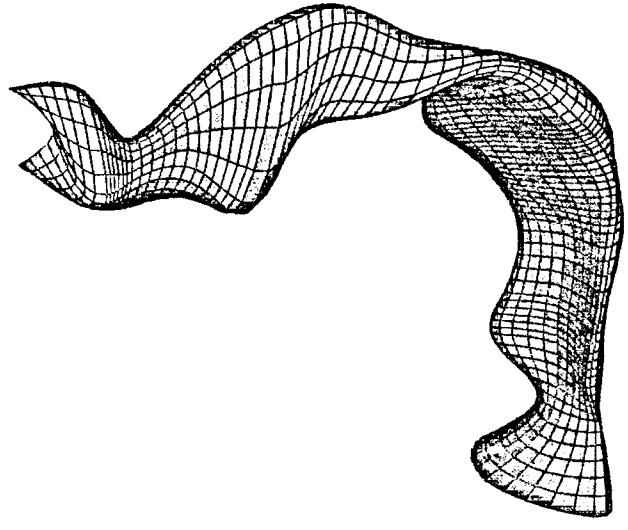


Figure 1. Lateral view of 3D model of the vocal tract tube for [ɑ]. The mouth is on the left. From an MR imaging based on 200 repetitions of the utterance [ɑi]. See [4].

in making ultrasound-based 3D models of the surface of the tongue for sounds where the tongue is in a central or back position. See figure 4.

Obviously ultrasound is a technique which offers tremendous possibilities for phonetic research. And it is only a matter of time before we see the first time-evolving three-dimensional models of the tongue based on ultrasound images with far better quality and shorter acquisition time than we use today.

During this 1/3 plenary session we intend to show examples of dynamic 3D models of the vocal tract.

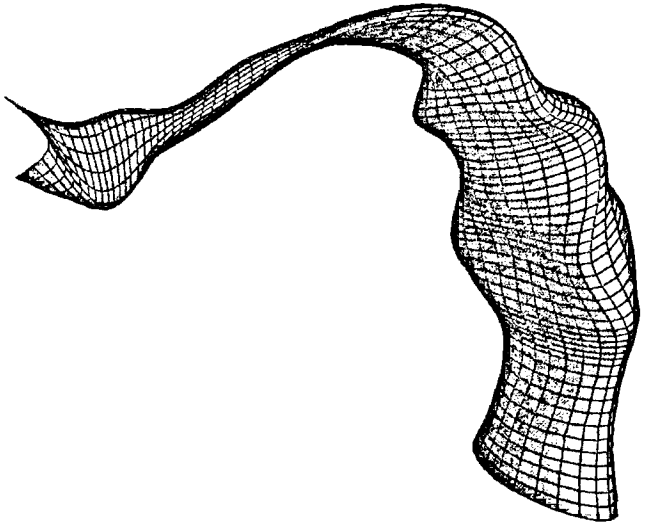


Figure 2. Lateral view of 3D model of the vocal tract tube for [i]. The mouth is on the left. From an MR imaging based on 200 repetitions of the utterance [ɑi]. See [4].

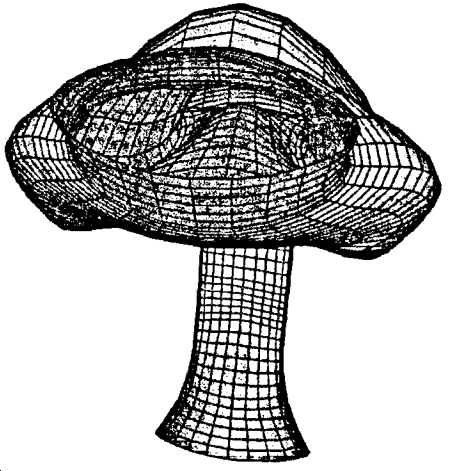


Figure 3. Frontal view of 3D model of the vocal tract tube for [ɑ].
From an MR imaging based on 200 repetitions of the utterance [ɑi]. See [4].

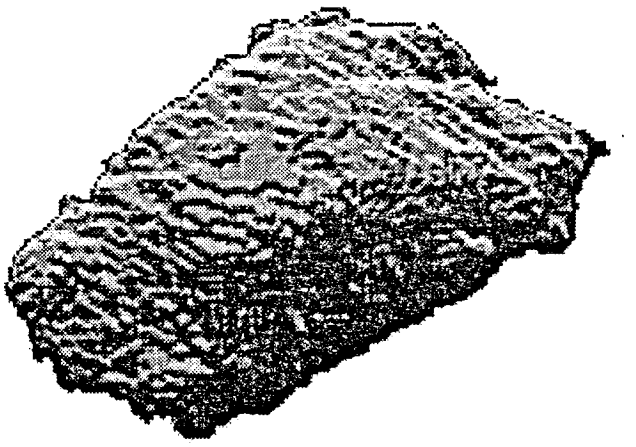


Figure 4. This figure shows a three-dimensional reconstruction of the surface of the tongue for [ɑ:]. Semi-lateral view of the tongue with apex to the left. The sulcus in the back part of the tongue can be seen. The ultrasound images were collected with a Vingmed Sound CFM-800 ultrasound scanner. The ultrasound probe was a mechanically steered annular array with 5 elements, producing a sector scan image with an opening angle of 90 degrees. The ultrasound frequency was 5 MHz. Resolution was 1 mm in the axial direction (along the ultrasound beam), and 3 mm in the transversal direction (normal to the ultrasound beam). The 2D image-frame rate was 35 frames/sec.The ultrasound probe was mounted in a custom made tilting device, using a stepper motor tilting the probe 0.7° per 2D image plane.

REFERENCES

[1] Rokkaku, M., Hashimoto, K., Imaizumi, S., Niimi, S. and Kiritani, S. (1986), "Measurements of the three-dimensional shape of the vocal tract based on the magnetic resonance imaging technique". *Annual Bulletin Research Institute of Logopedics and Phoniatrics*, vol. 20, pp.47-54.

[2] Yang, C.-S. and Kasuya, H. (1994), "Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects". *Proc. ICSLP 94*, pp. 623-626.

[3] Foldvik, A.K., Husby, O., Kværness, J., Nordli, I.C. and Rinck, P.A. (1990), "MRI (magnetic resonance imaging) film of articulatory movements.", *ICSLP90,* pp. 421-422.

[4] Foldvik, A.K., Kristiansen, U., Kværness, J. and de Bonnaventure, H. (1993), "A time-evolving three-dimensional vocal tract model by means of magnetic resonance imaging (MRI).", *Proceedings Eurospeech '93*, vol. 1, pp. 557 - 558.

[5] Grunmach, E. (1907), "Die Röntgentechnik zur Untersuchung der Mund-, Schlund- und Nasenhöhle bei der Phonation", *Archiv für Laryngologie*, vol. 19:3, pp. 405-407.

[6] Laver, J. (1994), *Principles of Phonetics*. Cambridge University Press.

[7] Stone, M. and Lundberg, A. (1994),"Tongue-palate interactions in consonants vs. vowels." *ICSLP94*, pp. 49-52.

DIFFERENCES AMONG SPEAKERS IN ARTICULATION OF AMERICAN ENGLISH /r/: AN X-RAY MICROBEAM STUDY

John R. Westbury[1], Michiko Hashi[1], and Mary J. Lindstrom[2]
[1]Waisman Center and Department of Communicative Disorders
[2]Department of Biostatistics
University of Wisconsin-Madison

ABSTRACT

Lingual fleshpoint positions and formant frequencies were measured at phonation onset in repetitions of the word *row* from an X-ray microbeam database including 55 normal speakers of American English. These data were used to develop a quantitative description of interspeaker variation in tongue "shape" for /r/, and to determine whether shape variations were acoustically significant, and/or related to gender and variation in selected measures of oral cavity size and shape.

INTRODUCTION

The x-ray microbeam (XRMB) technique is one of several contemporary methods for studying speech movement. The technique uses a narrow, high-energy x-ray beam (0.4 mm in diameter), controlled by computer, to track the real time motions of small gold pellets (2-3 mm diameter) glued to a speaker's head, lips, tongue, and lower jaw. Thus, the XRMB provides a *point-parameterized* view of speech movement [1], expressed in terms of the time-varying, digitally-sampled positions of discrete articulatory landmarks and fleshpoints.

The XRMB technique is not new. It has been available (albeit on a limited basis) for roughly 20 years, originally at the University of Tokyo where it was first implemented by members of the Research Institute of Logopedics and Phoniatrics [2]; and more recently, at the XRMB facility of the University of Wisconsin (UW) at Madison. The technique was developed as an alternative to high-speed, flood-field cineradiography, and has three significant advantages relative to that method, yielding more accurate data; involving significantly less exposure to ionizing radiation; and, imposing smaller data reduction burdens on the part of those who hope to analyze the information. A natural benefit is that it is now possible to collect and analyze data sets spanning many more speakers and task performances than were feasible using older methods. This "new" development, which we are just beginning to exploit, is important because many physical details of speech production behavior are variable within and across speakers. Generalizations about speech movement, for use in fields such as speech pathology, must be based on samples of speakers and tasks broad enough to reliably reflect the distribution of normal behaviors. Otherwise, there can be no good basis for distinguishing common, ordinary movements made by ordinary speakers, from those that are uncommon and extraordinary.

Over the past five years, a large-scale, freely-available speech production database has been developed at the UW XRMB facility. This database incorporates representations of lingual, labial, and mandibular movements, recorded in association with the sound pressure wave, for more than 50 normal, young adult speakers of American English, for a rich set of utterances and oral motor tasks, and lengthy recording interval (ca. 18 minutes/speaker). The large number of speakers makes this material especially well-suited for analyses of inter-speaker variation in articulatory kinematics.

We have selected a subset of materials from this database to examine production behavior for American English /r/. This sound is interesting and problematic from several points of view. From the acoustic theory of speech production [3], we know that the distinctively low third formant of /r/ can be approximated by vocal tract constrictions in three regions along the vocal tract length. This fact may be related to the kinds and frequency of misarticulations and substitutions that American children make for the sound, and also related to their tendency to master its production late in acquisition [4]. The sound /r/ is also unusual in the kind and degree of variation across major dialects of American English. Moreover, foreign speakers learning the language often find the American /r/ hard to say, and/or hear [5].

Some or all of these facts may be connected -- though precisely how is hard to say -- to an observation made by linguists for many years [6], to the effect that at least two distinct articulatory varieties of /r/ appear to exist, side by side. One broad type is the so-called *retroflex* variety, during which the tongue tip and blade are lifted, and the apex is curled backward in the mouth. The other is the *bunched* variety, where the apex and blade are held low while the front and dorsum of the tongue are elevated. In both varieties, speakers presumably attempt to achieve the same end, forming a primary oral constriction in the mid-palatal region. From a production point of view, /r/ is then interesting because there are different places along the vocal tract where its constriction can be formed, and for one of these places, different ways that the constriction can be formed.

Together, a handful of qualitative, descriptive studies [7-10] have addressed the basic accuracy of the retroflex-bunched distinction; and, only one of these, the remarkable study of Delattre and Freeman [7] of almost 30 years ago, attempted to describe variation in /r/ productions across a large speaker and task sample. We have set out to revisit the topic of /r/ production variability, among a speaker sample somewhat larger than that of Delattre and Freeman, and in so doing, address three specific goals. The first is to develop a quantitative description of variation in tongue posture or "shape" at an acoustically-defined *r-moment* in isolated examples of the word *row* produced by each speaker in the sample. The second goal is to determine whether variation across speakers in tongue shape, at a specific *r-moment,* might be related to variation in formant frequencies at (about) the same moment. And, the third goal is to determine whether the /r/ shapes assumed by speakers' tongues in *row* are related to gender, and/or to selected measures of oral cavity size and shape.

METHODS

The XRMB speech production database [11] incorporates material from 57 normal, native speakers of American English. Fifty-five (55) of these, 30 females and 25 males, are represented in our analysis of /r/. For this sub-sample, the median age was 21.0 years, with ages ranging between 18.3-37.0 years. For dialect purposes, 29/55 could be considered residents of Wisconsin, while 17 of the remaining 26 were residents of seven neighboring mid-western states of Minnesota, Illinois, Missouri, Iowa, Michigan, Indiana, and Ohio. Dialect homes of the remaining nine speakers were distributed across the US, from Massachusetts (1) to California (2).

Kinematic data recorded from each speaker represent the time-varying, mid-sagittally-projected positions of a set of articulator pellets. For 53/55 speakers, four such pellets were arrayed along the tongue midline. One of these (labelled T1) was always placed in the vicinity of the tongue blade, about 1 cm behind the apex of the extended tongue; a second (labelled T4) always placed in the vicinity of the tongue dorsum, about 6 cm behind the apex; and, two others (labelled T2 and T3) positioned to divide the interval between T1 and T4 into two roughly equal segments. For 2/55 speakers, only three tongue pellets were available. Other articulator pellets were attached to each speaker's mandible, and upper and lower lips.

Pellet-position data were expressed within a rectangular, anatomically-defined coordinate system [12]. The x-axis of the system corresponded to the intersection of

each speaker's midsagittal and maxillary occlusal planes. The y-axis was normal to the maxillary occlusal plane (MaxOP), and passed through a local origin at the point where the central maxillary incisors intersected that plane. Thus, *up* in this coordinate system points toward the top of the head along lines perpendicular to the MaxOP; and *forward*, toward the front of the face along lines parallel to the MaxOP, for each speaker.

Tongue shape measurements for /r/

Pellet positions were tracked at rates ranging between 40-160 times per second, as each speaker read through a list of records containing verbal and oral motor tasks. A subset of five records contained isolated instances of the word *row*, separated in time from different words in the same record, by 0.5-1.0 second. The moment of phonation onset was marked from oscillograms of the acoustic wave recorded during each instance of *row*, articulated by each speaker. Coordinates of all midline tongue pellets were extracted at the time of this event. These coordinates suggest the shape of the tongue at this discrete *r-moment*, and are the focus of our analyses.

In qualitative terms, most speakers prepare to say the /r/ of isolated *row* by drawing some forward part of the tongue up in the mouth, toward the palate, reaching an extreme local configuration some 50-100 ms before phonation onset. Speakers hold this posture for 50 ms or so beyond phonation onset, and then move the tongue rapidly downward, away from the palate, and variably forward or rearward, depending upon speaker and part of the tongue, toward a configuration suitable for the mid-back, diphthongal coda /o^{w}/.

Sample data from two speakers are shown in Figure 1. Shapes of the midline tongue contours at phonation onset for /r/ (computed across 4-5 repetitions of the word) are suggested by the average locations of pellets T1-4, connected by solid lines. Ensemble average pellet trajectories are also shown. These indicate paths traced by all four pellets during the interval spanning (-100,+500)ms relative to phonation onset. The pellet locations and trajectories are bounded above by piecewise continuous outlines of each speaker's palatal vault.

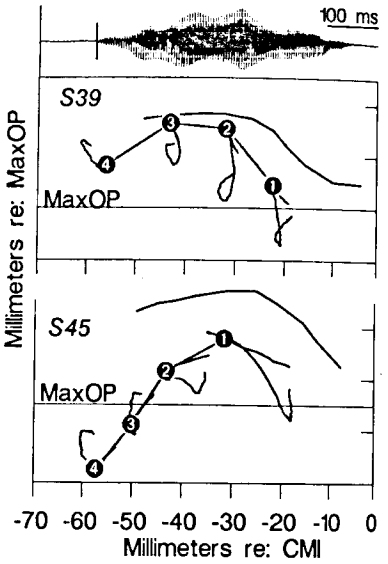


Figure 1. Above: Oscillogram of one subject's utterance of "row," and tongue contour at /r/ onset (marked by vertical line in the oscillogram). Below: Another subject's contrasting tongue contour.

Eight coordinates of four pellets provide a sense of tongue shape that is not very tractable. This is so partly because the number of values is high. A simpler expression of these data that reduces their dimensionality, and has the added advantage of emphasizing only tongue shape (and excluding tongue position) for each speaker, has the form of an ordered triple of angles, representing the orientation of straight lines drawn to connect positions of adjacent pairs of pellets. In our data, we designated the orientations of lines connecting pellet pairs {(T1,T2),(T2,T3),(T3,T4)} as angles (1,2,3), respectively. Angle triples (in degrees) for speakers 39 and 45, shown in Figure 1, were (-49,-5,32) and (28,56,50), respectively.

Acoustic measurements

Formant frequencies were measured from the digitized acoustic waveform, originally sampled at 21.74kHz. LPC and FFT spectra were generated using CSPEECH [13], for an analysis window 20 ms wide, centered at +20 ms with respect to phonation onset, for each token of *row* produced by each speaker. Estimates of formant frequencies from LPC analysis for each token were verified against wide-band spectrograms and corresponding FFT spectra. Bandwidths for spectrograms were 300Hz and 500Hz, for male and female speakers respectively, and the dynamic range was set to 72dB. The number of coefficients for LPC analysis was typically higher than the customary 24, and ranged between 30-40. The higher number of coefficients made certain formant identifications easier, and enhanced our ability to distinguish close second and third formants. Final acoustic measurements for each speaker represented mean formant values calculated across *row* repetitions.

Oral cavity size and shape

Three indices of oral cavity size and shape were derived from caliper measurements of stone models of each speaker's maxillary dental arch and palatal vault. These indices included mid-sagittal height of the palatal vault (*palht*), above MaxOP, measured 35 mm posterior to the central maxillary incisors; width of the maxillary arch (*m2wid*), measured between distal-buccal cusp tips of the second maxillary molars; and, distance rearward from the central maxillary incisors of the straight line connecting distal-buccal cusp tips of the second maxillary molars (*m2ap*), measured along a line parallel to MaxOP. A fourth index of cavity size -- distance from the central maxillary incisors rearward to the mid-sagittal outline of the posterior pharyngeal wall (*phap*), also measured along MaxOP -- was determined from a calibrated, sagittal-plane x-ray scan of each speaker's oral cavity.

Statistical methods

Several exploratory statistical techniques were used to gain insight into data collected for this study. All analyses were performed using S-Plus [14] or SAS [15]. The techniques included *hierarchial clustering* [16], to find transformations of the original pellet position data that captured intuitive shape information; *principal component analysis*, to determine the character and explanatory strength of various speaker-by-measure matrices; and *canonical correlation* [17] and *linear regression of ranks*, to search for associations between groups of measurements (e.g., between speaker-by-formant-frequency and speaker-by-tongue-shape matrices). The philosophy underlying data analysis was to capture and describe the extent and nature of variation among measurements of speakers and their articulations, in few terms, without greatly sacrificing interpretability.

Hierarchial clustering, which creates a hierarchy of groups from multi-dimensional data, played a central role in our attempts to describe and understand the pellet-position data. The result of an analysis of this type proceeds from one extreme where every individual is a group of one, to an opposite state in which all individuals form a single group. At each level within trees generated from such analyses, two groups which were closest together in terms of Euclidian distance were combined.

Statistical methods for choosing the "right" number of groups from data arrays do exist, but the methods are not robust. Moreover, they require two assumptions that we were unwilling to make. The first is that some underlying number of groups is already known to exist in the data. The second is that the data in each group follow some pre-specified distribution (e.g., multivariate normal). We chose not to look for some "right" number of clusters in our data, but to use entire hierarchical clustering trees to decide whether multivariate inputs to the

procedure (e.g., various transformations of the eight original pellet coordinates) captured shape information that was intuitively salient.

RESULTS

Tongue shape

Each speaker made essentially the same lingual gesture during /r/, achieving the same general tongue shape at phonation onset, each time they repeated the word *row*. When we view the entire set of (average) tongue shapes achieved by all speakers, we can readily point out some that look *bunched* (S39 in Figure 1), and others that look *retroflexed* (S45 in Figure 1). But, we also see shapes for some speakers that are not easily matched to these conventional descriptive labels. Two of these that are fairly easy to characterize are the tongue shapes that are relatively flat; and, those that are noticeably "tilted" (retroflexed?) but also somewhat convex (bunched?). Broadly, to the eye, these two shapes seem to be intermediate between those we might categorize as *retroflexed* and *bunched*, and therefore suggest a partition of the data into more than the two simple categories promoted by classical phonetic accounts. However, deciding how many categories there might be, and whether they correspond to the three suggested by Hagiwara [10], or the six suggested by Delattre and Freeman [7], or any other reasonable number (smaller than the speaker sample size!), is difficult. By eye, we can find speaker subsets amongst which roughly the same shape is achieved, though at the same time, it also seems true that these subjectively-identified subgroupings are never exhaustive or mutually exclusive.

More objective partitions, for a number of different expressions of the original pellet-position data, were obtained from hierarchical cluster analysis. However, the outcome of each such analysis was then judged subjectively, so that by eye, groupings found by the analysis had to be confirmed as "reasonable." The success of hierarchical clustering applied to tongue shape data was sensitive to the way the data were expressed. For example, different clusters, different numbers of major shape types, and different principal dimensions in the data were obtained from analyses on (1) the original speaker-by-coordinates array; (2) a similar speaker-mean-centered array (that removed tongue position from the original data); (3) a speaker-by-coefficients array (where the coefficients represented parameters of least-squares quadratic fits to the original coordinates); and, (4) the speaker-by-angles array that we finally found to be most useful. In part, we selected the angle expression of our original data because hierarchical clustering of the data array yielded groupings of speakers by tongue shapes that were intuitively satisfying. In Figure 2, we use scatterplots of tongue shapes to illustrate groups found from "angle" data. We have chosen to plot four groups because these give a visually pleasing result. However, we do not mean to imply that four is the "correct" number of groups.

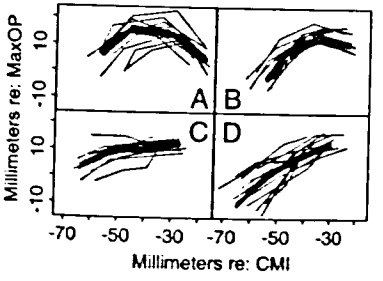


Figure 2. Scatterplots of tongue shapes showing four groups based on segment angles.

An angle expression of our data is also useful because it is easy to interpret in a way that directly suggests information about tongue shape. Speakers with negative first angles were those with a blade pellet (T1) that was lower in the mouth than the front-most intermediate pellet T2. Conversely, speakers with a positive first angle were those with the tongue blade (and T1 pellet) higher in the mouth, above MaxOp, than the portion of the tongue represented by the T2 pellet. The second angle was positive for speakers with strongly "tipped" (retroflexed?) tongues, and negative or near zero for those with more convex (bunched?) shapes. The third angle was positive for most speakers, indicating a dorsal pellet (T4) that was lower in the mouth (relative to MaxOP) than all other lingual pellets. The only speakers for whom angle three was not strongly positive were those with relatively flat tongue shapes.

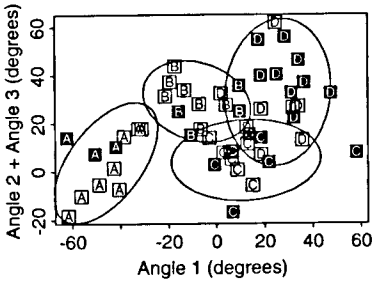


Figure 3. Tongue shapes plotted on an approximate principal component plane.

The first two principal components from an analysis of the speakers-by-angles data array explained 95% of the variability. This fact allows us to display tongue shape information, expressed in terms of angles, in two dimensions without a major loss of information. The first principal component was dominated by the first angle (defined by T1 and T2 pellet positions), while the second principal component was dominated by the sum of the second and third angles. The distribution of speakers' tongue shapes, expressed in terms of angles and plotted in the plane defined by these two approximate principal components, is illustrated in Figure 3. The approximate principal components are more easily interpreted than the actual components defined in analysis, and still explain 89% of the variability across speakers. Male and female speakers are distinguished by filled and open squares. Categories of four high-level shapes uncovered by hierarchical clustering are coded by letter, and two-sigma ellipses are drawn about coordinates of each category's centroid. Perhaps the simplest lesson that Figure 3 teaches is that the range of tongue postures for /r/, viewed in this way, is more nearly continuous than categorical across speakers, along either approximate principal component dimension. Simply put, this expression of our data seems to argue against discrete types of tongue shapes for /r/.

Formant frequencies

Table 1. Mean formant frequencies (in Hz) for male and female speakers. Standard deviations are in parentheses.

	Male	Female
F1	326 (39)	358 (45)
F2	882 (90)	1092 (122)
F3	1378 (121)	1792 (201)

Expected differences were found between formant frequencies for male and female talkers. Average frequencies for F1 and F2 agreed well with comparable data from normal geriatric speakers[18].

Interestingly, no significant association was found between speaker-by-tongue-shape and speaker-by-(log-transformed)-formant-frequency arrays. Thus, it appears that the very large differences across speakers in tongue shape at phonation onset in *row* do not seem to be accompanied by statistically reliable differences formant frequencies.

Oral cavity size

Table 2. Mean measures of oral cavity size (in mm), for males and females. Standard deviations are in parentheses.

	Male	Female
palht	22.4 (2.2)	18.7 (2.2)
m2wid	60.3 (4.7)	57.0 (3.1)
m2ap	43.7 (3.1)	41.5 (3.8)
phap	80.2 (5.0)	77.7 (4.1)

On average, males were slightly larger than females for all measures of oral cavity size. Male palates (*palht*) were about 3mm taller, and male maxillary arches at the second molar tooth (*m2wid*) were about 3 mm wider. Gender and oral cavity size are therefore confounded variables across our speaker sample. Across all speakers, the highest palatal vault was about 26 mm, while the shallowest was only 14 mm. The widest arch was 68 mm, while the narrowest was only 48 mm.

The only statistically significant relationship between our speaker-by-angle characterization of tongue shape for /r/, and measures of oral cavity size and shape, and gender, was between the first angle and gender. This effect is suggested in Figure 3, in which females, as a class, seem to have more negative first angles than do males. However, this effect was not strong. Gender explained only a small proportion of the variability in the first angle across speakers (r^2 = 0.13). Moreover, the association between the first angle (defined primarily by relative height of the tongue blade), and gender, has not been found in preliminary analyses of tongue shapes for /r/ in the words *street*, *problem*, *right*, and *across*.

DISCUSSION

How many kinds of tongue shapes exist for /r/ in American English? This simple question, asked by others before us, presumes that data should segregate into a number of discrete articulatory categories. However, our data seem to argue against such an assumption. No matter what visual and numeric tricks we have tried, the data summarized in this report do not seem to distribute well into discrete categories. It is probably closer to the truth that there is a continuous range of acceptable/possible tongue shapes for /r/. Speakers must achieve an acoustic result their listeners will accept as /r/. Precisely how that result is obtained, using the tongue and lips to constrict the vocal tract tube near either end, and/or near its middle, may be physiologically important to individuals, but not in a way that forces different speakers to achieve an invariant articulatory result. This is not a new idea, though tongue shape data for /r/, collected across many speakers, illustrate the idea perhaps more vividly than other data types.

The opportunity to examine data from many speakers is a main benefit of the XRMB method and database. Such an opportunity is necessary if we hope to know the distribution of tongue shapes for /r/ that exist in American English. This information is theoretically interesting, and may also have some practical benefit for speech therapy. Speakers who do not produce acceptable variants of /r/ are coached by therapists to achieve better results through instructions expressed in articulatory terms: instructions to shape the tongue in some particular way, and/or to place the tongue in some specific location. Speakers who are informed of the range of known and possible articulatory options may then choose some optimal variant.

The fact that variation across speakers in tongue shapes and formant frequencies at the same *r-moment* in *row* are not well related is superficially surprising. Differences between some speakers' tongue shapes, for the moment we have examined, are extreme. However, we can excuse the lack of relationship in view of standard acoustic theory [3]. The shape of the radiated spectrum depends heavily upon the vocal tract area function, and the area function itself is only partly determined by tongue shape within the oral cavity. The degree and locations of all constrictions along the vocal tract length define the area function. For /r/, constrictions in the pharynx, and at the lips, may be especially important. In our data, the former is inaccessible. The latter is somewhat less so, though the information available, given by positions of pellets on the lips, is difficult to interpret. Even the information we have for the oral portion of the tongue is less complete than we might like. Of course, the position of the apex is lost from our data, and this loss may be a special problem for any attempt to understand the articulation of /r/. We also have only a coarse outline of the tongue, defined by fleshpoints locations in the vicinity of the blade and dorsum, and two points in between. In principle, we might still expect some relationship between articulation and the acoustics of /r/, though for our data, we also have many reasons to reject that expectation.

The fact that variation across speakers in tongue shapes for /r/, and size and shape of the oral cavity, were not strongly related is also something of a surprise. We often assume that how speakers move when they speak depends partly upon how they are built. Our data for /r/ productions seem to show that this is not true, though it important again to emphasize the coarse nature of postural and size data. Certainly from our data, we cannot yet suggest why speakers choose the shapes they do for /r/.

ACKNOWLEDGEMENT

Research supported by USPHS Grant DC00820, and a collaborative research agreement between the University of Wisconsin-Madison, and ATR Human Information Processing Research Laboratories.

REFERENCES

[1] Houde, R. A. (1967), "A study of tongue body motion during selected speech sounds", Ph.D. dissertation, University of Michigan.

[2] Fujimura, O., Kiritani, S. & Ishida, H. (1973), "Computer-controlled radiography for observation of movements of articulatory and other human organs", Comput. Biol. Med., vol. 3, pp. 371-384.

[3] Fant, G. (1980), "The relations between area functions and the acoustic signal", Phonetica, vol. 37, pp.55-86.

[4] Shriberg, L. D. (1993), "Four new speech and prosody-voice measures for genetic research and other studies in developmental phonological disorders", Journal of Speech and Hearing Research, vol. 36, pp.105-140.

[5] Yamada, R. A. & Tohkura, Y. (1992), "Perception of American English /r/ and /l/ by native speakers of Japanese", in Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka eds, *Speech perception, production & linguistic structure*, Burke VA: IOS Press, Inc.

[6] Hockett, C. F. (1958), *A course in modern linguistics*, NY: Macmillan.

[7] Delattre, P. & Freeman, D. C. (1968), "A dialect study of American R's by X-ray motion picture", Linguistics, vol. 44, pp. 29-68.

[8] Zawadzki, P. A. & Kuehn, D. P. (1980), "A cineradiographic study of static and dynamic aspects of American English /r/", Phonetica, vol. 37, pp. 253-266.

[9] Lindau, M. (1985), "The story of /r/*", in V. A. Fromkin ed, *Phonetic linguistics, essays in honor of Peter Ladefoged*, NY: Academic Press.

[10] Hagiwara, R. (1994), "Three types of American /r/", UCLA working papers in phonetics, vol. 88, pp. 51-61.

[11] Westbury, J. R. (1994), *X-ray microbeam speech production database user's handbook, version 1.0*, Madison WI.

[12] Westbury, J. R. (1994), "On coordinate systems and the representation of articulatory movements", Journal of Acoustical Society of America, vol. 95, pp. 2271-2273.

[13] Milenkovic, P. & Read, C. (1992), *CSpeech Version 4 User's Manual*, Madison WI

[14] Becker, R. A., Chambers, J. M., Wilks, A. R. (1988), *The new S language: a programming environment for data analysis and graphics*, Belmont, CA: Wadworth.

[15] SAS Institute Inc. (1989), *SAS/STAT user's guide, version 6*, Carey, NC: SAS Institute Inc.

[16] Everitt, B. (1980), *Cluster analysis*, NY: Halsted.

[17] Dillon, W. R. & Goldstein, M. (1984), *Multivariate analysis, methods and applications*, NY: Wiley.

[18] Weismer, G., Kent, R. D., Hodge, M. & Martin, R. (1988), "The acoustic signature for intelligibility test words", Journal of Acoustical Society of America, vol. 84, pp. 1281-1291.

ELCTROMAGNETIC ARTICULOGRAPHY: A BRIEF OVERVIEW

Vincent L. Gracco
Haskins Laboratories, 270 Crown Street, New Haven, CT USA

INTRODUCTION

Up until the development of the first X-ray microbeam at the University of Tokyo the most common and reliable technique to transduce planar articulator motion during speech was cineradiography. As a result of the high radiation dosage the number and duration of such studies was necessarily limited. With the advent of the X-ray microbeam developed at the University of Tokyo [1] the ability to transduce speech articulatory motion with minimal radiation hazard allowed for longer experimental sessions. The expense to construct such a system, however, is considerable and the maintenance and operational costs are much too high to be a realistic consideration for any speech laboratory. Recently alternating magnetic field devices have become a less costly and more widely available alternative [2,3]. The utility of such devices will be outlined below as well as some of the considerations that will affect the sensitivity and reliabillity of the movement reproduction For a detailed evaluation of one of the devices currently in use and limited comparison to other commercially available devices, the interested reader is referred to Perkell, Cohen, Svirsky, Matthies, Garabieta, & Jackson [2].

There are currently three commercially available electromagnetic systems for the transduction of supraglottal articulation; the Carstens Electromagnetic Articulograph (EMA) [3], the Electromagnetic Midsagittal Articulometer (EMMA) [2], and the Movetrack [4]. The principles of operation are similar for all three devices. Transmitter coils, excited by a sinusoidal signal, produce an alternating magnetic field. A transducer coil, oriented parallel to the transmitter and transducer axes, will be induced with an alternating signal whose strength decreases appoximately in proportion to the cube of the distance from the transmitter. All current electromagnetic transduction devices for speech articulation research use monoaxial receiver coils placed on articulator flesh points in the midsagittal plane of the device. The signal induced in each receiver coil is the sum of the number of sinusoids (the number of transmitters) and the sampled voltages are subsequently processed to produce estimates of the distances between the receiver coil and the transmitter. The EMA and EMMA systems calculate the positions in software while the Movetrack system using a hardware implementation to estimate the locations of the receivers. Two of three commercially available systems (EMMA and EMA) use a three-transmitter design in which each transmitter is driven at three different carrier frequencies in the tens to hundreds of kilohertz range.

One of the benefits of the three-transmitter systems is the ability to correct for rotational misalignment between the transmitters and transducers. That is, any receiver misalignment, with respect to the magnetic lines of flux, reduces the estimated distance from the transmitter by the cosine of the misalignment angle. Both the EMA and EMMA systems provide methods for correcting for rotational misalignment while the Movetrack system, using a two-transmitter design, does not allow for any correction. A potential, and highly probable, error condition that neither the two- or three-transmitter systems is able to compensate for is off-midline displacement. Errors due to transducer misalignment with the midsagittal plane vary with position in the recording field [2,5]. Errors due to midline misalignment grow rapidly as a function of increasing distance from the origin of the device. During normal operation a combination of rotational misalignment and off-midline placement (0.5 mm for example) will result in errors ranging from .1 to 1.0 mm [2,5]. Because of the ability to correct for roatational misalignment within certain limits the most critical error factor is midline placement of the receivers in the midsagittal plane of the device.

It is also the case that the three-transmitter systems can operate at a lower field strength compared to the two-transmitter system [2]. For further information on the use of such systems for speech research the reader is referred to a recent publication resulting from an ACCOR workshop on the use of electromagnetic articulography in phonetic research [6].

APPLICATION

In this section a brief overview of the application of electromagnetic articulography to speech research will be considered. Data will be presented that have been collected using a version of the EMMA system [2] housed at Haskins Laboratories. The device has been operational since approximately 1992 following a series of tests to verify accuracy and reliability and to evaluate specific environmental influences [5]. Since the initial studies in our laboratory, the size of the transducers have been reduced by a factor of one-half resulting in receiver coils on the order of 2.5 x 2.5 x 1.0 mm.

A typical experiment consists of the placement of receivers on the bridge of the nose and the maxillary gum ridge (to monitor head motion during the experiment), receivers on the upper and lower lips, the mandibular gum ridge, and four locations on the tongue along with a sumultaneously recorded acoustic signal. At Haskins the nose, maxillary, mandibular, and lip receivers are attached using a biocompatible cyanoacrylate (Isodent). Attachment of the tongue receivers with Isodent requires extensive drying of the tongue surface and because the bonding is broken down by saliva, the attachment times can be quite short. As a result we routinely use Ketac bond to attach receivers to the tongue. . In contrast to Isodent the Ketac bond does not require the same degree of tongue surface drying and saliva has much less of an effect on the bonding of the surfaces. For a typical experiment tongue receivers have remained on the tongue surface for well over 90 minutes.

Signals from the EMMA, the acoustic signal, and any other simultaneously acquired signals (e.g., pressure, electromyographic, glottal transillumination) are digitized on-line (12 bit resolution) using a 64 channel A/D board. Our 10 channel EMMA system requires 30 input channels (3 voltages per channel) since the voltage-to-distance (V-D) conversion is done in software post-acquisition. Off-line calculations solve the near field equation obtaining x-y position of each receiver in the field at each point in time. Additionally, once head and occlusal coordinates have been established all data points are corrected for any head motion and rotated to the subjects occlusal plane.

Because the V-D conversion is done following digitization the voltages can be sampled at any rate that will be supported by the analog A/D board. However, this also means that any receiver problems during an experiment will not be apparent until after the experiment. To eliminate the possibility of wasting time and effort on collecting bad data, Perkell and colleagues at MIT have implemented a real-time display program that runs on a PC and is used to monitor receiver positions during an experiment.

An example of the two dimensional time history data obtained from receivers placed on four locations on the tongue is presented in Figure 1. Shown is a single repetition of a subject repeating the phrase "Say ladder again". The movement trajectories from the tongue have been digitally smoothed (23 point triangular window) following sampling at 625 Hz.

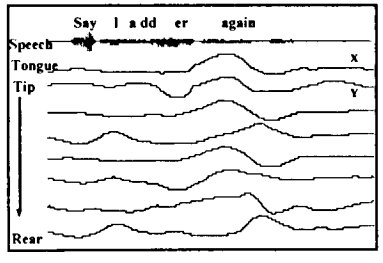


Figure 1. Tongue movement trajectories from four receivers equally spaced from the front of the tongue to the rear spanning a distance of approximately 4 cm.

Shown in the next figure is a single example of a subject repeating the phrase "Say tap again" displaying the motion of

the most anterior tongue receiver. At the bottom of the figure is the speed of the tongue tip associated with the two-dimensional trajectory. The dashed line illustrates the movement offset based on a minimum in the speed profile associated with the phonetic segment for /t/ in "tap".

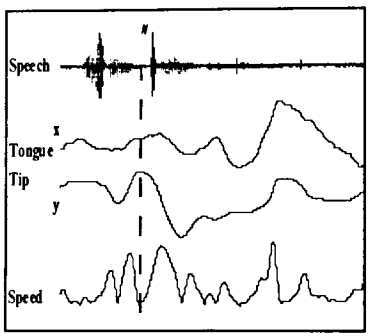


Figure 2. Tongue tip trajectory and the associated speed profile for the phrase "Say tap again". The /t/ closure is associated with the short duration steady state position observed in the tongue tip movement in the Y dimension.

From identification of motion onsets and offsets for a number of repetitions it is possible to acquire mean positions in two dimensional space of the tongue associated with a specific phone. Two dimensional coordinates can be obtained when the speed of the receiver is at a minimum within the acoustic duration of the phone of interest. The data are then fitted using a cubic spline interpolation and an estimate of tongue shape obtained. Figure 3 reflects the average of ten repetitions in which tongue shape was estimated for /r/ in the word "rack" and "heard". Comparing the tongue shape using a cubic spline interpolation with actual shapes obtained from static midsagittal magnetic resonance images have been generally good.

The acquired data can also be displayed as receiver paths in the sagittal plane. Figure 4 is the same data from Figure 1 displayed in this manner in which the form of the articulator paths can be easily visualized. Also presented in the figure is an outline of the hard palate taken during the experimental session. It is also possible to estimate the constriction degree (in the midsagittal plane) by simply locating the minimum distance of a receiver from the palate location at the time of minimum speed .

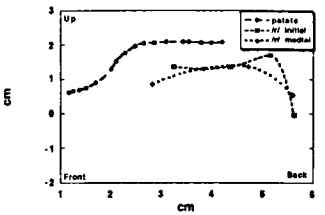


Figure 3. Estimated tongue shape for /r/ when produced word intial ("rack") and syllabic ("heard"). Each points represents the average of ten repetitions with spatial locations obtained at the time of minimum speed for each receiver.

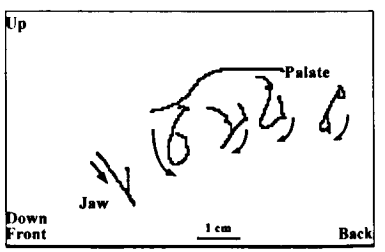


Figure 4. Tongue and jaw paths in occlusal space associated with the word "ladder". Arrows indicate the direction of the motion.

CONCLUSION

Electromagnetic articulography is fast becoming an important tool in the acquisition of large quantities of speech articulation data. Provided that a number of precautions are taken, the precision that can be acheived by such devices can be quite high. While the total costs are often considerable, taking into account the required hardware and software, the acqusition of such devices are within the financial reach of many institutions. Moreover, the use of electromagnetic articulography to clinical populations may provide important breakthroughs in understanding a variety of speech movement disorders [see 7 for example]. With increased use further refinements will be forthcoming in both hardware and software.

ACKNOWLEDGMENT

This work was supported by Grants DC-00121 and DC-00594 from the National Institute on Deafness and Other Communication Disorders.

REFERENCES

[1] Fujimura, O, Kiritani, S., & Ishida, H. (1973). Computer-controlled radiography for observations of articulatory and other human organs. *Comp. Biol. Med*, 3, 371-384.

[2] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992), "Electro-magnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements", *J. Acoust. Soc. Am.*, vol. 92, pp. 3078-3096

[3] Schönle, P., Gräbe, K., Wenig, P., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31, 26-35.

[4] Branderud, P. (1985). Movetrack-a movement tracking system. *Proceedings of the French-Swedish Symposium on Speech*, GALF, Grenoble, France, 113-122.

[5] Gracco, V. L., & Nye, P. W. (1993). Magnetometry in speech articulation research: some misadventures on the road to enlightenment. *Proceedings of the ACCOR Workshop on Electromagnetic Articulography in Phonetic Research*Institute fur Phonetik und Sprachliche Kommunikation der Universitat Munchen, 31, 91-104.

[6] *Proceedings of the ACCOR Workshop on Electromagnetic Articulography in Phonetic Research.* (1993). Institute fur Phonetik und Sprachliche Kommunikation der Universitat Munchen.

[7] Van Lieshout, P., Alfonso, P., Hulstijn, W., & Peters, H. (1993). Electromagnetic articulography (EMA) in stuttering research. *Proceedings of the ACCOR Workshop on Electromagnetic Articulography in Phonetic Research.* Institute fur Phonetik und Sprachliche Kommunikation der Universitat Munchen, 31, 215, 224.

RECENT ADVANCES IN HIGH-SPEED DIGITAL IMAGE RECORDING OF VOCAL CORD VIBRATION

Shigeru Kiritani
Research Institute of Logopedics and Phoniatrics, University of Tokyo, Tokyo, Japan

ABSTRACT

The paper describes recent dvances in the high-speed digital image recording of vocal cord vibration. One is the introduction of a large size image memory which enables data recording of longer duration and thus, observations of involuntary, sporadic changes occuring in certain pathological voices. Another is the development of the system of higher frame rate (4500frames per second with 256 x 256 pixels). Examples of the data analysis conducted by these sytems are presented.

INTRODUCTION

To study voice source characteristics in speech, it is important to record vocal cord vibration simultaneously with the speech signal and to analyze the relationship between the pattern of the vocal cord vibration and the acoustic characteristics of the speech signal. The system of high-speed digital image recording developed by the present authors is convenienet for this kind of studies and the system has been used at our institute for the studies of voice source characteristics in normal speech as well as in pathological voices[1-3].

In order to further fascilitate such studies, several technical improvements were introduced to our original system. The present paper reports on the recent advances in our system.; one is the use of a large size image memory for recording glottal image in longer phonation, and another is a development of a new system with higher frame rate and higher resolution. Characteristics of the improved system together with the examples of the data obtained by this system will be presented below.

Fig. 1 shows a block diagram of the high-speed digital image recording system. The syste consists of an oblique angled solid endoscope, a camera body containing an image sensor, and a digital image memory. The laryngeal image obtained through the endoscope is focused on the image sensor. The image sensor is scanned at a high frame rate and

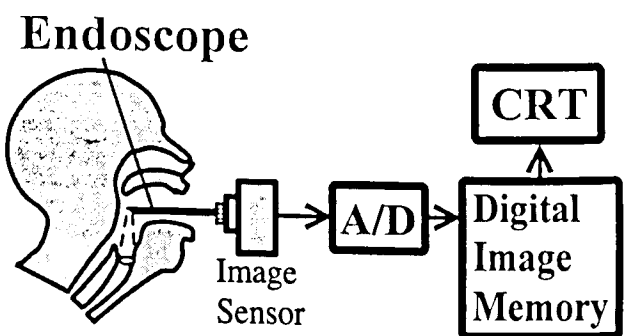


Figure 1 Blockdiagram of high-speed digital image recording system.

the output video signal is fed into the image memory through a high-speed A/D converter. Stored images are then reproduced consecutively as a slow-speed motion picture. In our original system, we used a commercially available image sensor. In order to achieve a high frame rate, it was necessary to scan only a selected part of the sensor. The frame rate was 2500 per second with the number of pixels 126 x 32.

LARGE-SIZE IMAGE MEMORY -Image Recording of Longer Duration-

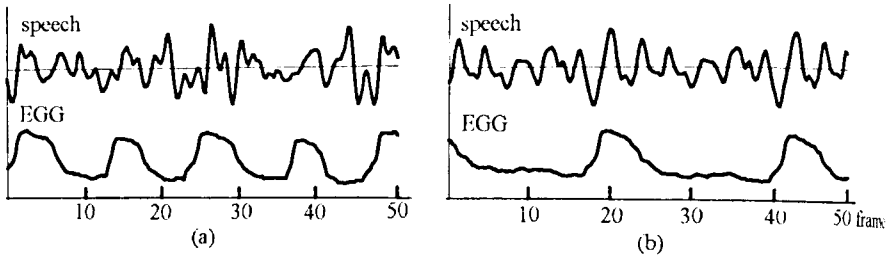
High-Speed observation of vocal cord vibration has generally been conducted for very short periods during sustained phonation, typically a fraction of a second. However, there are several kinds of studies which require data recording of longer duration. One example is the analysis of vocal cord vibration during running speech which includes consonants. For this purpose, a high-speed digital image recording system combined with a flexible fiberscope is now being used at our laboratory. This kind of study requires the recording of laryngeal behavior in the natural utterances for duration of a few seconds.

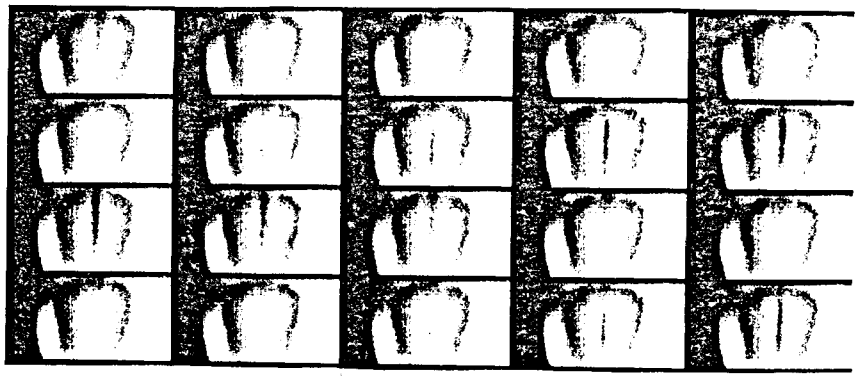
Another example is the analysis of vocal vord vibration associated with sporadic, involuntary voice changes in certain pathological cases. In such studies, recordings of several seconds duration are desired to catch the moments of sporadic changes in the vocal cord vibration such as changes in the pitch frequency.

In order to carry out such studies, a special, large-size dididal image memory was constructed. The size of the memory is 64 Mbyte, and it can store 15,000 frames of glottal images with 126 × 32 picture elements. This corresponds to an image recording of 6 seconds at a rate of 2,500 frames per second. Below, two examples of pathological vocal cord vibration are presented which are associated with sporadic, involuntary changes in the fundamental period of the voice occurring during sustained phonations.

Case 1 shows a rough voice accompanied by sporadic changes in the fundamental period. The subject does not show any apparent pathological change in his vocal cords and he underwent a botulinus toxin injection 4 months prior to the recording for treatment of his spasmodic dysphonia. Figures 2 (a) and (b) show the speech and EGG signals of his voice during the periods of normal pitch and lowered pitch in the same phonation. The pitch period for the lowered pitch is nearly the twice that for the normal pitch. The glottal image during the period of normal pitch shows a clear, tight closed phase in each vibratory cycle. In contrast, the glottal image during the period of lowered pitch shows that the glottal closure is incomplete in every other vibratory cycles, there is no apparent movement of the EGG signal or excitation pattern in the speech wave in these cycles. Thus, it can be concluded that in this phonation, weakening of the closing movements of the vocal cords in every other cycles brings about the apparent doubling of the fundamental period in the speech signal. An a Apparent fundamental period in speech and EGG signals actually corresponds to 2 cycles of vocal cord vibration.

The voice of case 2 is characterized by intermittent cessations in voicing and is accompanied by marked changes in voice quality as well as in pitch frequency. Figures 4 (a) and (b) show the speech and EGG signals for his





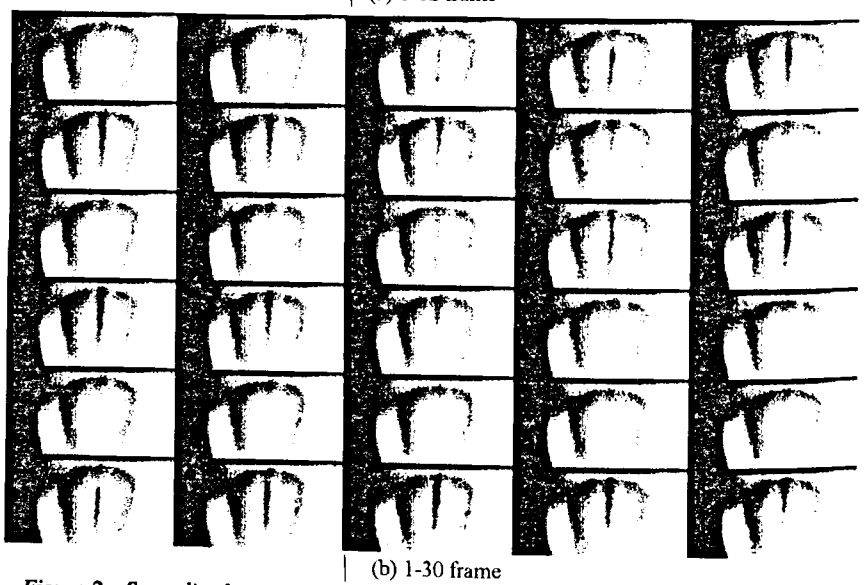
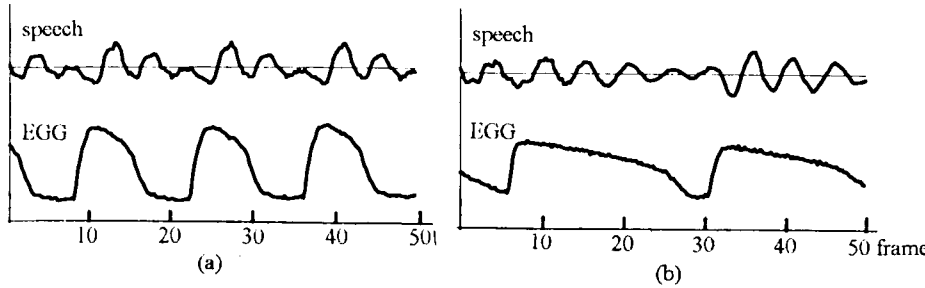


Figure 2 Sporadic change in the fundamantal period of voice; case 1. (a) Period of normal pitch. (b) Period of lowedred pitch. Frame rate 2500/second.



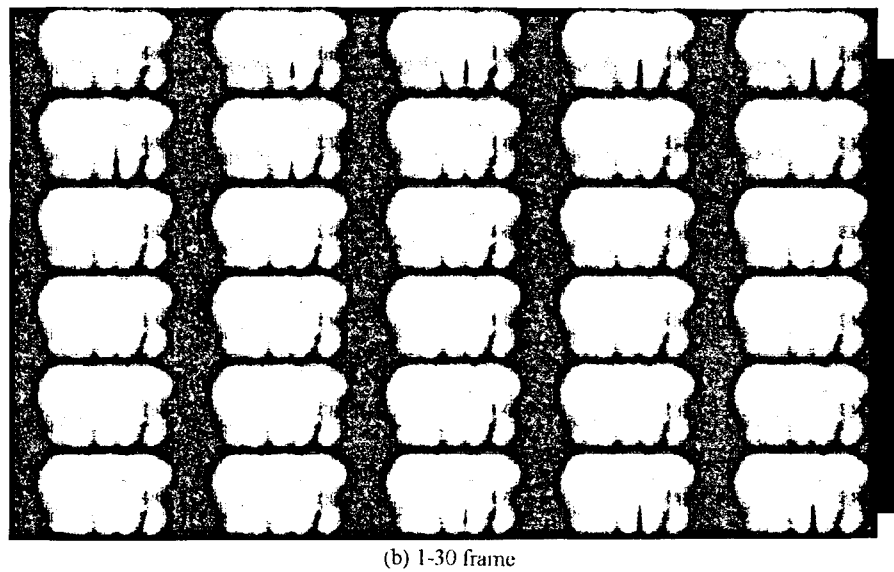


Figure 3 Sporadic change in the fundamantal period of voice; case 2. (a) Period of normal pitch. (b) Period of lowedred pitch. Frame rate 2500/second.

voice during the periods of normal pitch and lowered pitch. In this case also, the fundamental period of the speech wave and the EGG signal for the lowered pitch are about twice those for normal pitch. However, the glottal images confirms that the vocal cord vibration during the period of lowered pitch has a long period of glottal closure which is accompnied by the short period of glottal opening. In this case, an elongation of the fundamental period of the speechwave results from the longer closure period in the vocal cord vibration.

Thus, in the present study, both the voices of case 1 and case 2 show involuntary, sporadic changes in the fundamental period of the speech wave during sustained phonation. In the period of lowered pitch, the pitch period is nearly twice that in normal pitch. However, high-speed recording of the

glottal image revealed a characteristic difference between the 2 cases. In case 1, it was due to the weakening of the closing movement in every other vibratory cycle, and in case 2, it was due to a true elongation of the vibratory cycle.

A NEW HIGH-SPEED, HIGH-RESOLUTION SYSTEM

As described above, in our original system, the maximum frame rate was limited to 2,500 frames per second with 126×32 pixels. Recently, a new system with a higher frame rate and higher resolution was developed in cooperation with Photron Co. Ltd. Photron Co. Ltd has produced a specially designed image sensor which incooperates a technique of parallel read-out of image signals to obtain a high frame rate. The sensor contains 256×256 picture elements and can be scanned at a rate of 4,500 frames per second. When the image area is restricted to 256×128 picture elements, the frame rate is 9,000 per second. As an example of data analysis obtained by the new system, an analysis of vocal cord vibration in a simulated diplophonic voice is presented below. The voice was produced by a normal subject simulating a diplophonia.

Figure 4 shows the speech signal in this phonation. The speech signal shows quasi-periodic variations in amplitude and waveform in 9 pitch periods. A waveform with large amplitude and strong excitation is observed in every 9th period. In between these cycles, the speech amplitude gradually gets smaller. In our previous paper, we reported on the chatacteristics of vocal cord vibration in pathological cases of diplophonia[4]. It was observed that there is a difference in vibratory frequency between the left and right vocal folds, and that the phase difference between the vocal cords varies with time quasi-periodically, resulting in a quasi-periodic variation in the speech signal. However, in that study, due to limitations in image resolution, it was difficult to identify the edges of the vocal cords and to measure the movements of the vocal cords from the glottal images, and only a qualitative measure of the glottal width was presented.

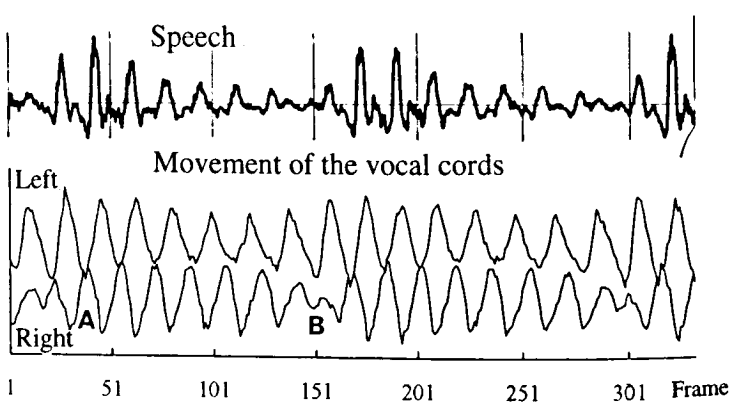


Figure 4 Movements of the vocal cords in a simulated diplophonic voice produced by a normal subject. Frame rate 4500/second.

In the present data, owing to the improved image quatily, the edges of the vocal cords can be determined by a visual inspection of the glottal image, at least for this phonation and therefore more detailed information on the movements of the vocal cords can be obtained. Figure 4 (c) shows the result of such measurement. Movement of the edges of the left anf right vocal cords were measured on the selected horizontal scan line shown on the glottal image in the figure. It can be seen in the figure that at around A in the figure, the movements of the vocal cords are nearly in phase, and that the glottis shows a period of complete closure. Then, during successive cycles, the phase difference becomes progressively larger. At around B in the figure,the inward movement of the right vocal fold is incomplete, and one vibratory cycle of the right vocal cord almost disappears, and this process cancels and resets the phase difference between the left and right vocal cords.

The temporal change in the pattern of the vocal fold vibration described above explain the pattern of the temporal change in the speech waveform.

SUMMARY

Recent technical advance in high-speed digital image recording of vocal cord vibration were prresented. Preliminary experiments cerfirmed that data recording of longer duration is useful for observing sporadic phenomena occurring during sustained phonations in certain pathological cases. A new system of higher frame rate made possible more quantitative analysis of movements of vocal cords.

REFERENCES

[1] K. Honda, S. Kiritani, H. Imagawa and H. Hirose: "High-speed Digital Recording of Vocal Fold Vibration Using a Solid-state Image Sensor," in Laryngeal Function in Phonation and Respiration, College-Hill Publication, 485-491, 1987.

[2] H. Imagawa, S. Kiritani and H. Hirose: "High-speed Digital Image Recording System for Observing Vocal Fold Vibration Using an Image Sensor," J. Medical Electronics and Biological Engineering, 25, 284-290, 1987.

[3] S. Kiritani, H. Imagawa and H. Hirose: "Vocal Cord Vibration and Voice Source Characteristics--- Observations by the High-speed Digital Image Recording," Proc. ICSLP-90, Kobe, Japan, 61-64, 1990.

[4] S. Kiritani, H. Imagawa and H. Hirose: "High-speed Digital Image Analysis of Vocal Corld Vibration in Diplophonia," Speech Communication, 13, 23-32, 1993.

COMPUTATIONAL PHONETICS

Roger K. Moore
DRA Speech Research Unit, Malvern, UK

ABSTRACT

Recent impressive advances in the capabilities of systems for automatic speech recognition and automatic speech generation has meant that there is a growing need to unify the emerging theoretical and practical developments in speech technology with the established knowledge and practices in the phonetic sciences. This trigger paper discusses some of the relevant issues and proposes the establishment of a new discipline to be known as *Computational Phonetics*.

BACKGROUND

The idea that an automatic device can be configured to 'recognise' or 'synthesise' human speech not only has the practical benefit of providing human operators with hands-free eyes-free control of equipment or access to information, but could also be said to provide the ultimate test for phonetic theories of human speech perception and production.

Moreover, it is precisely in the area of 'speech technology' (particularly in automatic speech recognition and automatic speech generation) that the experimental and descriptive fields of phonetics, linguistics and psychology meet the computational disciplines of artificial intelligence, computer science and engineering.

Both of these observations raise interesting issues concerning the role of contemporary phonetics in the light of the substantial advances that are currently being made in the capabilities of automatic speech recognition and generation systems.

Automatic Speech Recognition

Automatic speech recognition has come a long way from the first simple attempts back in the 1950s. In the early days, vocabularies were small (usually the ten digits), the words had to be uttered in 'isolation' (that is, with a distinct pause between each word) in a quiet environment, and each user was obliged to 'train' the system by providing a set of example utterances - whole-word 'templates' - against which subsequent words to be recognised would be compared (thereby rendering the process 'speaker dependent').

Forty years on, automatic speech recognition systems can operate with vocabularies containing many thousands of words, the input can be natural 'continuous' speech and, after estimating the parameters of a set of suitable statistical models (for example, 'hidden Markov models' - HMMs) using data from an appropriate spoken language corpus, utterances can be recognised from a wide range of 'independent' speakers operating in more natural environments (such as in an office or over a telephone).

Automatic Speech Generation

Likewise, speech synthesis systems have progressed from manually operated electrical and mechanical devices to automatic text-to-speech reading machines which can be adapted to exhibit the vocal characteristics of a desired target speaker and which can handle abbreviations and acronyms as well as regular textual input.

Also the process of speech generation from first principles using a mathematical analogue of the human production apparatus is being supplemented by an approach based on the concatenation of relevant fragments of natural human utterances which have been extracted from an appropriate spoken language corpus using automatic processes not dissimilar to those used in automatic speech recognition.

Prerequisites for Progress

One might easily imagine that these substantial advances have been caused by the implementation of linguistic and phonetic 'knowledge' in such speech technology systems. However, it can be argued (particularly for automatic speech recognition) that progress has in reality been a direct result of the introduction of rigorous mathematical and statistical modelling paradigms coupled with the development of efficient 'search' and 'parameter estimation' algorithms supported by a phenomenal increase in available computing power and data handling capacity.

It can also be argued that further progress depends on a continued concerted effort to tackle some of the theoretical and practical issues in automatic speech recognition and generation, not the least of which is to arrive at a greater understanding of the structure and regularities of speech signals themselves and of the 'process' which relates an audio-visual speech 'pattern' to it's cognitive counterpart. Such an understanding might be expressed in terms of a *theory* of 'speech pattern processing' [1].

SPEECH PATTERN PROCESSING

Speech essentially mediates the expression and communication of ideas, concepts and information between different physical entities through a regularity of behaviour which is shared, and hence 'understood', by the participants. It is this regularity of behaviour - the *patterning* - which is central to speech pattern processing and hence to speech recognition and generation. It is the patterning which provides the 'constraint' which allows human behaviour never before encountered to be recognised and interpreted appropriately, and which conditions the generation of novel behaviour never before required.

Speech Patterning

Information about the patterning in speech is derived from two principal sources; (i) the discipline of phonetics (and related areas such as psycho-acoustics, linguistics, psycho-linguistics etc.) which provides descriptive 'knowledge' about the observed regularities in speech, and (ii) annotated speech corpora which provide hard 'evidence' for more detailed speech pattern behaviour.

Thus far, neither source of constraint is sufficient on its own to facilitate high-accuracy automatic speech recognition and generation. However, it is fair to say that it is the extensive use of large-scale speech corpora that has been the key to the success of current automatic speech recognition and generation systems.

Of course it is not sufficient simply to have (even detailed) information about the constraints implicit in speech patterning in order to construct a functional automatic speech recogniser or synthesiser; it is also necessary to define a (set of) 'representation(s)' with which to 'encode' such constraints.

Likewise, the appropriateness of any given representation depends critically on the 'computation' which is to be performed upon it - and such an 'algorithm' needs to be founded on some kind of mathematical 'theory' of recognition or generation.

Speech Pattern Processing Theory

Thus far, the most successful approaches to automatic speech recognition have been based on the theory of 'maximum-likelihood' (or Bayes') classification which defines the interpretation of a sequence of acoustic observations in terms of the most probable explanation taken over all possible interpretations. From this theory it is possible to derive a mathematical and statistical 'modelling' paradigm (such as hidden Markov models) which provides a suitable integrated representation of acoustic, phonetic and lexical constraints together with compatible algorithms for estimating the model parameters from annotated data and for computing the most likely interpretation of an unknown input sequence.

On the other hand, automatic speech generation is founded on less well developed formalisms and, as such, lags behind recognition in it's theoretical sophistication. Low-level processes such as the generation of a spectrum from a parametric representation of a vocal tract are based on solid mathematical principles, but the control of such parameters is often handled in a more heuristic manner. However, the introduction of statistical techniques (more familiar to automatic speech recognition) for control parameter modelling is beginning to take place.

Stochastic Modelling

It is important to appreciate that the use of statistics in speech pattern processing is convenient simply because it provides a rigorous mathematical framework for modelling 'uncertainty' and for characterising the processes of 'approximation', 'interpolation' and 'extrapolation' which are all key components of the requirement to be able to categorise unseen data and to be able to generate novel data.

The value of stochastic models in general, and HMMs in particular, is that the formalism shows no signs of being limited in the extent to which it can be developed to accommodate more complex modelling requirements; the mathematics has already been extended to handle simultaneous asynchronous events (thereby removing the 'single synchronous signal' assumption) and to include dynamic segmental effects (thereby removing the frame-to-frame 'independence' assumption).

Both of these advances point towards a possible unification of HMM structures with the modelling strategies normally employed in speech synthesis and the ideas expounded in the field of 'non-linear phonology' [2]. However, this unification can only be achieved if there is effective communication between the appropriate specialist practitioners involved in the speech pattern modelling and phonetics areas.

THE ROLE OF PHONETICS

Clearly, in principle, the field of phonetics has a great deal to contribute to the design of appropriate annotated speech corpora and to the expression of the phonetic and linguistic 'priors' which might be made implicit in a system's modelling structures.

However, both of these activities must be carried out in full cognisance of the theoretical and mathematical implications involved; *it is not appropriate to propose new representations without considering whether they are compatible with any known scheme for computation.*

It is therefore proposed that the skills and expertise represented by the phonetic science community could be usefully directed *not* towards the construction of better automatic speech recognisers or synthesisers, but towards the exploitation of the theoretical and practical tools and techniques from speech technology for the creation of more advanced theories of speech perception and production (by humans *and* by machines). Indeed it is perhaps now appropriate to begin to think in terms of establishing a new more balanced discipline which could be described as *'Computational Phonetics'*.

Practitioners in this new area should be encouraged to work towards a *unified* theory of speech pattern processing which could answer some of the outstanding fundamental questions about speech [3] to the benefit of *both* speech technology and speech science.

REFERENCES

[1] Moore, R. K. (1993), "Whither a theory of speech pattern processing", Proc. EUROSPEECH'93, pp 43-47.

[2] Moore, R. K. (1994), "Speech pattern processing: from 'blue sky' ideas to a unified theory?", Proc. UK Inst. of Acoustics Conf. on Speech and Hearing, pp 1-13.

[3] Moore, R. K. (1994), "Twenty things we still don't know about speech", Proc. CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology.

THREE AREAS OF COMPUTATIONAL PHONETICS

Hans G. Tillmann
Institute of Phonetics and Speech Communication, University of Munich

ABSTRACT

We propose to consider three areas of computational phonetics. The first area deals with speech signals transmitting phonetic information from the effectors to the receptors of natural nervous systems. The second one deals with the categories of phonetic facts accessed in large speech databases. The third one is devoted to computing time functions of given phonetic categories and to predicting the categories evoked by phonetic time functions.

INTRODUCTION

The term "computational phonetics" could be given many meanings. Is it the brother (or sister) of Computational Linguistics? This would restrict the meaning of the term to those parts of phonetics which make use of computers in a specific way. Indeed, many examples of this type of computational phonetics could be described, since computers, equipped with analog-to-digital and digital-to-analog converters, have become the most important, or even, in a number of cases, the only instrument of instrumental phonetics.

We also could look back into the history of instrumental phonetics where we will find clear instances which likewise could fall under the term computational phonetics. There are at least two prominent examples which we would like to mention.

The first one is probably the earliest instance of computational phonetics. We find it in the appendix of Scripture's "Elements of Experimental Phonetics" where the author, nearly 100 years ago, showed how the Fast Fourier Transform of a voiced speech signal can be computed using paper and pencil, after the amplitude of a pitch-period from a graphically recorded oscillogram has been optically magnified and manually sampled into equidistant discrete amplitude values.

Another very important historical example which must be mentioned here was presented in 1957 in an article entitled "Die Vokalartikulation als Eigenwertproblem" by Meyer-Eppler and Ungeheuer in Zsch. f. Phonetik. The authors made use a second order homogeneous differential equation (Webster's Horn equation) and showed how the three-dimensional geometry of the vocal tract can be reduced to the two-dimensional area function in order to compute the resonance frequencies of the human vocal tract during the production of vowels. Thus the values of formant frequencies became uniquely predictable by a mathematically formulated physical theory (cf. also Ungeheuer 1957, 1962 and Fant 1960).

It seems to make sense to restrict the term computational phonetics to theories that take values of some given type and compute new values which are not trivially available in the computed form. This restriction is useful only to exclude what shall not fall under a computational theory. So we need further criteria to determine which types of phonetic theories shall be part of computational phonetics. In the following we would like to argue that there are exactly three interesting areas of computational phonetics which should be particularly considered and metatheoretically investigated in more detail. It could well be the case that the future of phonetics as one of the speech sciences will depend on further theoretical developments in just these three areas.

(I) THE AREA OF CAUSAL SPEECH THEORIES: THE TRANSMISSION OF PHONETIC INFORMATION

During any real speech act many physical processes which remain transphenomenal to the normal speakers and listeners can be measured and represented in the form of digitally sampled discrete time functions. As soon as a speaker conducts an act of speech these physical processes are assumed to run in the brains of speakers and listeners as well as in all channels that connect the involved individual nervous systems; they synchronously accompany any naturally perceived speech utterance.

These processes are extremely complex time functions inside the verbally communicating neural systems and still widely unknown with respect to their specific segmental and prosodic neural form. However, at the very periphery of the communicating systems they become fairly simple [15] and can be easily represented by well manageable AD-converted discrete time functions. This is the reason why causal theories of speech transmission allow us to fill the gap between the communicating neural systems. The articulatory CVCVC-actions of the motor system of the speaker cause speech movements which are discretely mapped to the proximal receptors of the speaking system and are also indirectly transformed into the acoustic output which transmits all relevant phonetic information to the auditory receptors (also the visible speech input plays a role).

The theory of Meyer-Eppler and Ungeheuer has been further developed by Schroeder (1967) and others; so it is no problem today to compute the acoustic output from a given articulatory time function digitally representing the speech movements of a given speech utterance [1,12]. If we know the impulse response of the acoustic system and the given excitation, the output can be simply derived by a convolution (in the time domain) or by a multiplication of the z-Transforms (in the frequency domain). Thus in the first area of computational phonetics the transmission of phonetic information is made explainable by showing how the peripheral actions of the motor system cause a mathematical mapping to the receptors of the sensory systems of speakers and listeners.

(II) THE AREA OF CATEGORICAL SPEECH THEORIES: THE REPRESENTATION OF PHONETIC FACTS

The world of phonetic facts that are relevant to the speakers and listeners of a language can be satisfyingly represented by logically oriented programming languages (such as Prolog). They allow computation with these facts in a very effective way as soon as there is access to a database large enough to contain all possible instances of those facts. And if the utterances in the database are represented according to the CRIL conventions of the IPA (concerning the Computer Representation of Individual Languages [4]), the lexical items they are composed of can not only be identified by their orthographic form, but also automatically compared as to their canonic citation forms and the actual and factual realisations (varying in a Lindblomian H-H-space [9]).

This type of computationally approach to phonetic facts has been systematically developed by Christoph Draxler in the German Verbmobil-PhonDat-project [2]. The experiences up to now show two things: Categorical representation of phonetic facts on the 3rd CRIL-level, aligned to individual

sound segments, is still somewhat critical in the case of reduced words in spontaneous speech [3]. Here, new standards for representing less clear and unclear cases have to be established (IPA and SAMPA etc. have been mainly used for description of well articulated clear speech utterances). Secondly, it should be pointed out that symbolic data receive a new kind of factuality if they are connected to real speech utterances within a large database (even if the physically recorded speech signals are not analysed themselves, but only used for the instantiation of the given facts). This condition allows us to say that Prolog is used to automatically analyse real phonetic facts by taking nothing but categorical representations from databases as input and so to compute new phonetic knowledge.

(III)
THE AREA OF EXTENSIONAL SPEECH THEORIES: FROM SYMBOL TO SIGNAL, FROM SIGNAL TO SYMBOL

The values of Prolog variables and predicates are not restricted to symbolically represented phonetic facts, but can also be extended to relate the possible categories and the analysable physical properties of the speech signal to each other, in both directions. There is of course no analytical relation between symbols that represent categorical facts and speech signals which *per se* are nothing but digitized time functions. In a speech database we empirically identify both sides according to Feigl's principle [15,6]: any complex category can be experimentally reproduced by repeating a given time function, and the time functions that we record when a category is repeatedly demonstrated (by different speakers in different situations) will be again instantiations of that category. Thus we can look for *aposteriorily necessary* connections (in the sense of Kripke's (1980) analysis) in order to answer the question as to what properties of the signal cause the perception of a category, and which properties are to be expected if the category is reproduced within a speech act. This will ultimately allow us to define the physical extensions of the phonetic categories of a spoken language.

In the PhonDat-project it has been demonstrated by Florian Schiel that it is a good step in this direction to compute the phonetic facts on a yet unspecified 3rd CRIL-level, by using the information of the 2nd canonic level as input to a speech verification procedure [19].

However, the final aim of this third area of C.P. is to be able to compute any speech signal that falls in the extension of a given category, and to determine the phonetic form that the words of a spoken language segmentally and prosodically take as soon as they are used by the speaker of this language in connected speech.

REFERENCES

[1] Carré, R. and M Mrayati (1990), *Articulatory-acoustic-phonetic relations and model-ling, regions and modes.*, pp. 211-240 in: W. Hardcast-le and A. Marchal (eds.), Speech pro-duction and speech modelling. Dordrecht.

[2] Draxler, C. (1995), *Intro-duction to the Verbmobil-PhonDat Database of Spoken German*, Prolog Applications Conference, Paris.

[3] Eisen, B., H. G. Tillmann, and C. Draxler(1992), *Consistency of judgements in manually labelling of phonetic segments: The distinction between clear and unclear cases,* ICSLP Banff pp. 871-875.

[4] Esling, J. (1990), *Computer-coding of the IPA: Supplemen-tary Report*, JIPA 20, pp.

[5] Fant, G. (196l), *Acoustic theory of speech production*, The Hague: Mouton & Co.

[6] Feigl, H. (1958), *The 'mental' and the 'physical'*, pp.370-497 in Feigl et al.(eds): Minnesota studies in the philosophy of science, Vol. II, Minneapolis.

[7] Kohler, K (1990), *Segmental reduction in connected speech in German: phonological facts and phonetic explanations*, pp. 69-92 in: W. Hardcast-le and A. Marchal (eds.), Speech production and speech modelling. Dordrecht.

[8] Kripke, S. A. (1980), *Naming and necessity*, Cambridge: Harvard University Press

[9] Lindblom, B. (1990), *Explaining phonetic variation: A scetch of the H and H theory* pp. 403-439 in: W. Hardcastle and A. Marchal (eds.), Speech production and speech modelling. Dordrecht.

[10] Meyer-Eppler, W., und G. Ungeheuer (1957), *Die Vokalartikulation als Eigenwertproblem,* Ztschr. f. Phonetik 10, pp. 245-257.

[11] Moore, R.: *Twenty things we still don't know about speech*, pp. 9-7 in H. Niemann, de Mori and Hanrieder (eds.): Progress and prospects of speech research and technology, Sankt Augustin 1994.

[12] Rabiner, L. R., and R. W. Schafer (1978), *Digital proces-sing of speech signals*, Englewood Cliffs: Prentice-Hall.

[13] Schiel, F. (1995), *An automatic segmentation program based on HMM*, internal report, to appear in FIPKM.

[14] Schroeder, M. (1967), *Determination of the geometry of the human vocal tract by acousic measurements*. Jasa 41, pp.1002-1010.

[15] Tillmann, H. G., (1993), *Why articulation matters in SLP*, FIPKM 31, pp. 11-28.

[16] Tillmann, H. G., (1995), *Kleine Phonetik und Grosse Phonetik*, to appear in Kohler-festschrift, Phonetica.

[17] Ungeheuer, G.(1957), *Untersuchungen zur Vokalartikula-tion*, Phil. Diss., Bonn.

[18] Ungeheuer, G. (1962), *Ele-mente einer akustischen Theo-rie der Vokalartikulation*, Ber-lin: Springer.

[19] Wesenick, B., and F. Schiel (1994), *Applying speech verification to a large database of German to optain a statistical survey about rules of pronunciation,* pp. 279-282, ICSLP, Yokohama.

PHONETICS OF SECOND-LANGUAGE ACQUISITION: PAST, PRESENT, FUTURE

Winifred Strange
Communication Sciences & Disorders
University of South Florida
Tampa, FL, USA

ABSTRACT

This paper summarizes the recent history of research on three issues in second-language (L2) phonetics: a) predicting the relative perceptual difficulty of L2 phonetic categories, b) describing the relationship between perception and production of L2 phones, and c) optimizing perceptual training to improve perception of L2 phonetic contrasts.

INTRODUCTION

The following three papers of this semi-plenary session report on three areas of research on the perceptual phonetics of second-language (L2) acquisition which have had a long and productive history. In this paper, I summarize the empirical progress on these topics over the last 25 years and discuss briefly how the theoretical and methodological issues have evolved.

THE PAST

The field is indebted to Arthur S. Abramson & Lee Lisker, who reported their seminal findings on cross-language differences in the perception of voice onset time (VOT) at the VIth ICPhS [1]. That study demonstrated that the ability to *discriminate* differences in VOT which underlie voicing and aspiration contrasts in initial stop consonants in many languages was predictable from language-specific patterns of phonetic labeling of the synthetic stimulus continuum. This finding fit well with the theoretical claims of Motor Theory [2] that the perception of speech sounds was accomplished via special processes that were intimately related to speech production. The *categorical perception* (CP) paradigm, which compared performance on tests of (physical identity) discrimination and (phonemic) identification of synthetically-generated acoustic continua, provided a rigorous methodological tool to examine these language-specific patterns of perception. Problems in perceiving non-native vowels were largely ignored in early cross-language research because (synthetic, steady-state) vowels were not perceived categorically and did not show language-specific patterns of perception [3].

Cross-language CP studies with adults in the 1970s replicated and extended the finding that discrimination of acoustic continua underlying voicing and place contrasts among consonants was determined by the phonemic significance of the stimuli in the listeners' native language [4,5]. At the same time, developmental research demonstrated that 2- to 6-month-old infants *could* discriminate place and voicing contrasts in consonants, whether or not they had been exposed to a language in which the contrasts occurred [6,7,8]. Thus, it was concluded that there was some *loss* in discrimination ability as a function of age or learning one's native-language phonology, or both.

This conclusion was reinforced by two additional kinds of data on the perception of non-native consonant contrasts by adults. One finding, which was first reported by Goto [9] see also [10], demonstrated that native Japanese speakers of English who had learned to produce /r/ and /l/ correctly nevertheless failed to distinguish these liquids perceptually when presented their own or native English speakers' productions, i.e., production preceded and exceeded (auditory) perception in L2 learning. Second, early training studies (using synthetic stimuli) which attempted to improve the perception of non-native contrasts met with limited success [11, 12,13]. While subjects' performance improved on training materials, generalization to new tasks and novel stimuli, including natural speech utterances, was limited. Thus, adult L2 learners' perceptual problems appeared to be serious and very long lasting, if not permanent. These results provided empirical support for the strong Critical Period Hypothesis proposed by Lenneberg [14].

In retrospect, these conclusions, based on a very limited number of phonetic contrasts, experimental paradigms, and subject groups, were premature, overstated, and in some respects, incorrect. Cross-language research in the 1980-1990s, which expanded the investigation to additional contrasts and subject groups using new stimulus materials and testing techniques, improved our understanding of the phenomena in all three areas of research.

Relative Perceptual Difficulty

Questions about the perceptual difficulty of an extended set of non-native contrasts were explored using carefully constructed natural speech materials (as well as synthetic stimuli) and an expanded variety of perceptual tests. For instance, Gottfried [15] demonstrated that both monolingual English speakers and experienced L2 learners of French had difficulty perceptually differentiating French front rounded vowels in a *categorial* (name identity) discrimination task (see Jamieson's paper). English listeners also had difficulty distinguishing French /e-ɛ/, which constituted a native phonemic contrast but whose members differed in phonetic detail (see also Bohn's paper). Werker and Tees [16,17] reported that adult English listeners had more difficulty categorizing a non-native place contrast than a non-native voicing contrast in Hindi stops; difficulty with the place contrast persisted even after one year of Hindi instruction [18]. Polka [19] further demonstrated that the Hindi place contrast differed in perceptual difficulty (for English listeners) as a function of the voicing context in which it occurred. Best and her colleagues [20] reported that both voicing and place contrasts among Zulu clicks were well discriminated by native English speakers, despite their being unlike any native phonetic categories. Thus, (initial) difficulty in perceiving both consonant and vowel contrasts ranged from minimal to extensive.

Experiments on the effects of L2 experience suggested that perceptual differentiation of non-native contrasts improved with immersion experience or intensive conversational instruction [21]. However, perception of some contrasts did not reach native-like levels even after years of experience. Furthermore, experiments using synthetic speech in which multiple acoustic cues for a contrast were manipulated independently indicated that L2 learners based their perceptual responses on different acoustic cues than native listeners. For instance, relatively inexperienced Japanese L2 learners of English appear to base their perceptual differentiation of (syllable-initial) /r-l/ more on temporal differences and on F2 spectral cues, than on the F3 spectral cue that is considered the primary differentiating parameter for native listeners [22,23]. Flege [24] reported that inexperienced Arabic learners of English assigned more perceptual weight to vowel duration than to consonant duration cues for voicing contrasts in final fricatives, whereas more experienced learners showed a native-like trading relation.

Developmental cross-language research continued to produce significant insights

with regard to the ontogeny of language-specific patterns of perception. Werker and her colleagues [25] published some remarkable experiments demonstrating that between 6 and 12 months of age, English-learning infants showed a decline in their ability to differentiate non-native Hindi and Salish consonant contrasts. More recently, Polka & Werker [26] reported the emergence of language-specific patterns of vowel perception at an even earlier age. It thus appears that native-language patterns of phonetic perception are formed in the first year of life. In addition, one study suggests that exposure to an L2 before the age of 2 years old has lasting consequences for later perceptual learning [18].

The data are not consistent concerning whether children between the ages of 2 and 13 years old have any advantage over adolescents and adults in the perception of non-native contrasts. Flege and Eefting [27] found that many (but not all) Puerto Rican children who started learning English at the age of 5-6 years had English-like perceptual boundaries on a VOT continuum, whereas older learners displayed perceptual boundaries that were a compromise between native Spanish and native English locations. However, other studies [28, 29, 16] failed to show better perception of non-native contrasts by preadolescent L2 learners.

Perception/Production Relationships

Although it is often assumed that perceptual difficulties lead to incorrect or accented production of non-native phonetic categories by L2 learners, until quite recently, there have been few studies that directly assess the relationship between perception and production in L2 learning (see Llisterri's paper). Rochet [30] demonstrated that the perceptual assimilation patterns of speakers of different languages are predictive of L1 substitution patterns in the production of French /y/. Portuguese speakers assimilated the non-native /y/ to their /i/ category while native English speakers assimilate the same stimuli to their /u/ category. These differences in perception accounted for production patterns, suggesting a causal relationship between perception and production of L2 phonetic categories, at least in the early stages of L2 acquisition. However, recent research by Yamada and her colleagues has shown that perception and production may proceed independently in L2 learning. In a study of a large group of Japanese learners of English with different amounts of immersion experience, perception generally lagged behind production such that perception mastery was a good predictor of production mastery, but the reverse relation did not hold.

With respect to questions about the effects of age-of-learning on L2 phonetics, it has been well documented that "earlier is better" with respect to learning to produce non-native phonetic segments with little or no accent. However, as mentioned earlier, the same advantage has not been demonstrated convincingly for L2 perception.

Perceptual Training of L2 Contrasts

Studies conducted in the 1980-90s have demonstrated that short-term intensive training can improve perception of non-native consonant contrasts when the appropriate stimuli and tasks are employed (see Jamieson's paper). Non-native voicing contrasts appear to be easier to learn than place contrasts [18,31]. However, Pisoni and his colleagues [32] have demonstrated that Japanese performed significantly better on the difficult /r-l/ contrast after completing 15 sessions of identification training with a large corpus of natural speech minimal pairs. Yamada [33] further demonstrated that performance continued to improve over 45 training sessions, and for some subjects, reached native-like levels of performance.

An interesting finding of recent training studies is the significant role that syllable context plays in limiting the extent of generalization. Morosan & Jamieson [34] reported that while training native French speakers on the English /ð-θ/ contrast in synthetic CV syllables improved perceptual differentiation of natural speech CV utterances, there was no significant transfer to the contrast in VCV or VC contexts. Apparently, subjects learn to differentiate position-specific allophones of phonetic categories, rather than context-free phoneme categories.

THE PRESENT

Results of research in the 1980-90s increased our understanding of the phenomena of L2 phonetics. Several conclusions can now be drawn, although many questions remain unanswered. (See [35] for further reviews.)

Conclusions from Recent Research

1) Both children and adult L2 learners have significant difficulties perceptually differentiating some, but not all, vowels and consonants that are not functionally distinctive in their native language. They may also have difficulty differentiating phonetic categories that are phonemic in their native language, but differ in their phonetic realization in the L2.

a) These perceptual difficulties are *not* due to a loss of sensory capabilities, but rather reflect *perceptual attunement* to phonetic information that is phonologically relevant in the native language. Language-specific patterns of selective perception are formed very early in L1 acquisition.

b) Since all non-native phonetic contrasts are not equally difficult; contrastive analysis of phoneme inventories cannot accurately predict perceptual problems of L2 learners. Perceptual difficulty varies as a function of the phonotactic and phonetic context in which the non-native contrasts occur. It may be that temporally cued contrasts are easier to perceive than spectrally-cued contrasts (but see below).

2) selective perceptual patterns are modified in adult L2 learners (as well as children) through immersion in the L2 environment or intensive conversational instruction. Perception of L2 contrasts may continue to improve for several years. However, some perceptual difficulties may persist, even after production of non-native phonetic segments is mastered. Thus, while L1 substitution patterns in production by inexperienced L2 learners are predictable from perceptual assimilation patterns, perception and production mastery may be uncorrelated in more experienced L2 learners.

3) Short-term training using stimuli and tasks that emphasize *equivalence classification* (rather than discrimination of physical differences) can lead to significant and lasting improvement in the perception of non-native contrasts. Such training has been shown to transfer to novel talkers and stimuli (i.e., new phonetic contexts) but, to date, generalization across different phonotactic contexts has not been demonstrated.

Current Theories of L2 Phonetics

Current research on the phonetics of L2 learning focuses on several remaining questions about the nature of the language-specific patterns of perception, the relationship between L2 perception and production, and the effects of perceptual training on L2 perception (and production) patterns. While good descriptive studies are still being conducted (and provide very valuable data), more current experimentation is theory-driven. Current theoretical debates center on some basic questions regarding how to characterize phonetic categories, L1 and L2 categorization processes, and *what* is learned during perceptual training.

Two working models have been offered that attempt to predict (and explain) the relative perceptual difficulty of non-

native phonetic categories. They complement each other in that Best's Perceptual Assimilation Model (PAM) [36] focuses on initial perceptual difficulties, while Flege's Speech Learning Model (SLM) [37] proposes an account of perceptual reorganization in both L2 and L1 as a function of L2 experience.

According to PAM, non-native phonetic segments are perceptually assimilated to native phonetic categories according to their articulatory-phonetic (gestural) similarity to native *gestural constellations*. If the non-native phones are very discrepant from any native phonetic gestures, they may be assimilated as uncategorizable speech or even as non-speech sounds. Perceptual difficulty in differentiating a non-native contrast is predictable from these assimilation patterns. If the contrasting phones are both assimilated as good exemplars of a single native category, perceptual differentiation is extremely difficult; if the contrasting phones differ in their "goodness of fit" to a single native category, then perception will be somewhat easier. If the two phones are assimilated to two different native categories, they will be differentiated with ease. Finally, non-assimilated phones will be perceptually differentiated on the basis of their psychoacoustic distinctiveness.

In a recent version of his SLM, Flege proposes that L1 and L2 *position-sensitive allophones* are related along a continuum of interlingual phonetic similarity, defined in acoustic-phonetic terms, such as the F1/F2 formant space for vowels or the VOT continuum for voicing in stop consonants. He hypothesizes that beginning L2 learners perceptually assimilate most L2 segments to native categories; however, if the L2 segment is sufficiently dissimilar from any L1 segment that L2 learners can discern the difference perceptually, then a new L2 perceptual category will be established over time. For less dissimilar L2 segments, separate L2 category formation may continue to be blocked because of equivalence classification of L1 and L2 segments. In these cases, a single perceptual category subsumes both L1 and L2 segments, leading to persistent accented production of the L2 segment and even to shifts in production of the native segment away from the monolingual norm.

Although these two models differ in the emphasis placed on acoustic vs. articulatory specification of phonetic similarity, they both take context-dependent phonetic segments as the appropriate level of analysis rather than the more abstract phonemes or distinctive features of traditional linguistic analysis.

Other theorists are concerned with the nature of phonetic category organization and how it affects the perception of native and non-native phonetic segments. Pisoni [38] argues that an exemplar-based model can best account for several phenomena in speech perception, including why training with a large and variable corpus is successful in reorganizing phonetic categories. According to this model, perceivers store detailed information about individual phonetic segments, including speaker-specific and context-specific information. Categorization involves matching an incoming signal on the basis of its overall physical similarity to previously stored exemplars. Thus, native phonetic categories are represented as clusters of exemplars that share certain critical (acoustic) parameters, while varying on other, non-criterial characteristics. For L2 perceptual learning to be successful, training must be conducive to the formation of (new) equivalence clusters.

In Kuhl's *Perceptual Magnet* model [39] native-language phonetic categories are organized around best cases or *prototypes* (established within the first year of life) which distort the phonetic perceptual space. Acoustic variations around these prototypes come to be perceived as more similar to each other. This warping of the perceptual space around native-language prototypes accounts for the failure to differentiate phonetic variants that are distinctive in the L2 but constitute within-category variations in the L1. L2 perceptual learning would require the reorganization of the phonetic perceptual space around newly established prototypes.

Each of these four theorists makes somewhat different claims about the nature of the stored representations of L1 phonetic categories. However, all depend on one or another definition of *phonetic similarity* as an organizing schema. An important task for future research is to characterize the notion of phonetic similarity in explicit and non-circular ways.

THE FUTURE

Research on the phonetics of L2 acquisition is a vibrant and productive area of endeavor. While there have been great advances in our understanding of the basic phenomena in the last 10 years, important unanswered questions about the very nature of phonetic categories and categorization processes involved in the perception of speech remain. In my remaining comments, a few suggestions for future research are made.

1) In describing assimilation patterns in L2 perception, it is important that experiments be conducted with stimuli and tasks that tap perceptual processes at appropriate levels of analysis. Thus, when investigating the relative salience of temporal cues *vs* spectral cues, it must be remembered that both kinds of information are imbedded in a larger context in continuous speech. Temporal cues for phonetic contrasts are defined relative to other gestural timing characteristics (stress, rate of speech) which are specified over larger stretches of speech than single syllables. Spectral cues also vary as a function of coarticulation and timing patterns. So, for instance, similarity of L1 and L2 vowels ascertained from judgments of vowels produced in isolation or citation-form syllables may not predict perceptual assimilation patterns in more naturally produced utterances [43]. Further, the phonotactic contexts in which phonetic contrasts are investigated influence the results profoundly. Language-specific knowledge of allophonic variation and syllable structure rules interacts with (language-universal) constraints to determine listeners' expectations about how phonetic segments influence each other in speech utterances. There is a need for research that investigates how this aspect of L1 phonology affects L2 perception/production patterns.

2) Research on L2 perception and production suggests that their interrelationship may change in complex ways over a relatively long period of time. More research is needed which traces these changes over sufficiently long time periods. Because studies of L2 perception and production show large individual differences, long-term longitudinal studies are needed.

3) Perceptual training studies have concentrated primarily on L2 consonant contrasts where members of the contrasting pair sound "the same." On the other hand, L2 vowel contrasts are usually discriminable even from the outset. Thus, the perceptual problem is one of learning which of the discriminable differences are critical for the contrast, and which others constitute within-category variations. With respect to production/perception relationships, vowels and consonants may also differ. Whereas consonant gestures involve contact of articulators (with concommitant tactile feedback), vowel articulation requires spatial positioning of the tongue in a relatively open vocal tract. It may be the case, therefore, that production of vowels is more dependent on auditory feedback. Training studies

that assess effects of perceptual training on L2 vowels and consonants will provide important insights into these differences. [Work supported by NIDCD]

REFERENCES

[1] Abramson, A.S. & Lisker, L. (1970) "Discriminability along the voicing continuum: Cross-language tests." *Proc. VIth ICPhS* 569-573.

[2] Liberman, A. M., et al. (1967). "Perception of the speech code." *Psychological Review*, 74, 431-461.

[3] Stevens, K.N., et al. (1969) "Cross-language study of vowel perception." *Lang. Speech* 12, 1-23.

[4] Miyawaki, K., et al. (1975) "An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English." *Percept. Psychophys.*, 18, 331-340.

[5] Williams, L. (1977) "The perception of stop consonant voicing by Spanish-English bilinguals." *Percept. Psychophys.*, 21, 289-297.

[6] Lasky, R. E., et al. (1975) "VOT discrimination by four to six and a half month old infants from Spanish environments." *J. Exper. Child Psychol.*, 20, 215-225.

[7] Streeter, L. A. (1976) "Language perception of two-month old infants shows effects of both innate mechanisms and experience." *Nature*, 259, 39-41.

[8] Trehub, S.E. (1976) "The discrimination of foreign speech contrasts by infants and adults." *Child Development*, 47, 466-472.

[9] Goto, H. (1971). "Auditory perception by normal Japanese adults of the sounds "L" and "R"." *Neuropsychologia*, 9, 317-323.

[10] Sheldon, A., & Strange, W. (1982) "The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception." *Applied Psycholing.*, 3, 243-261.

[11] Lisker, L, (1970) "On learning a new contrast." *Haskins Laboratories : SRSR*, SR-24, 1-17.

[12] Strange, W. (1972) "The effects of training on the perception of synthetic speech sounds: Voice onset time." Unpubl. Ph.D. dissertation, U. Minnesota.

[13] Strange, W. & Dittmann, S. (1984) "Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English." *Percept. & Psychophys.*, 36, 131-145.

[14] Lenneberg, E. (1967) *Biological foundations of language*. NY: Wiley.

[15] Gottfried, T.L. (1984) "Effects of consonant context on the perception of French vowels." *J. Phonetics*, 12, 91-114.

[16] Werker, J.F., & Tees, R.C. (1983). "Developmental changes across childhood in the perception of non-native speech sounds." *Canadian J. Psychol.*, 37, 278-286.

[17] Werker, J.F., & Tees, R.C. (1984) "Phonemic and phonetic factors in adult cross-language speech perception." *J. Acoust. Soc. Am.*, 75, 1866-1878.

[18] Tees, R.C. & Werker, J.F. (1984) "Perceptual flexibility: Maintenance or recovery of ability to discriminate non-native speech sounds." *Canad. J. Psychol.* 38, 579-590.

[19] Polka, L. (1991). "Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions." *J. Acoust. Soc. Am.*, 89, 2961-2977.

[20] Best, C.T., et al. (1988) "Examination of perceptual reorganization for non-native speech contrasts: Zulu click discrimination by English-speaking adults and infants." *J. Exp. Psychol.: Human Percept. Perform.*, 14, 345-360.

[21] MacKain, K.S., et al. (1981) "Categorical perception of English /r/ and /l/ by Japanese bilinguals." *Applied Psycholing.*, 2, 369-390.

[22] Underbakke, M., et al. (1988) "Trading relations in the perception of /r/-/l/ by Japanese learners of English." *J. Acoustic. Soc. Am.*, 84, 90-100.

[23] Yamada, R. A., & Tohkura, Y. (1992) "Perception of American English /r/ and /l/ by native speakers of Japanese." In Y. Tohkura, E. et al. (Eds.) *Speech perception, production and linguistic structure*. Tokyo, JAPAN: OHM Publishing Co. Ltd. 155-174.

[24] Flege, J.E. (1984) "The effect of linguistic experience on Arabs' perception of the English /s/ vs. /z/ contrast." *Folia Linguist.*, 18, 117-138.

[25] Werker, J.F. & Tees, R.C. (1984) Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant Behav. Develop., 7, 49-63.

[26] Polka, L. & Werker, J.F. (1994) "Developmental changes in the perception of nonnative vowel contrasts." *J. Exper. Psychol.: Human Percept. Perform.*, 20, 421-435.

[27] Flege, J.E. & Eefting, W. (1987) "Production and perception of English stops by native Spanish speakers." *J. Phonology*, 15, 67-83.

[28] Snow, C.E., & Hoefnagel-Hohle, M. (1977) "Age differences in the pronunciation of foreign sounds." *Lang. Speech* 20, 357-365.

[29] Cochrane, R.M. (1980) "The acquisition of /r/ and /l/ by Japanese children and adults learning English as a second language." *J. Multiling. Multicultural Develop.*, 1, 331-360.

[30] Rochet, B.L. (in press) "Perception and production of L2 speech sounds by adults." in [35].

[31] McClaskey, C.L., et al. (1983) "Transfer of training of a new linguistic contrast in voicing." *Percept. Psychophys.*, 34, 323-330.

[32] Logan, J.S., et al. (1991) "Training Japanese listeners to identify English /r/ and /l/: A first report." *J. Acoustic. Soc. Am.* 89, 874-886.

[33] Yamada, R.A. (1993) "Effect of extending training on /r/ and /l/ identification by native speakers of Japanese." *J. Acoustic. Soc. Am.* 93, 2391.

[34] Morosan, D.E., & Jamieson, D.C. (1989) "Evaluation of a technique for training new speech contrasts: Generalization across voices, but not word position or task." *J. Speech Hear. Res.* 32, 501-511.

[35] Strange, W. (ed.) (in press). *Speech perception and linguistic experience: Issues in cross-language speech research.* Timonium MD: York Press.

[36] Best, C. (in press) "A direct realist view of cross-language speech perception." in [35].

[37] Flege, J.E. (1991). "Speech learning in a second language," in Ferguson, D. et al. (eds.) *Phonological development: Models, research, and application.* Parkton, MD: York Press.

[38] Pisoni, D.B. & Lively, S.E. (in press) "Variability and invariance in speech perception: A new look at some old problems in perceptual learning." in [35].

[39] Kuhl, P.K. & Iverson, P. (in press) "Linguistic experience and the "perceptual magnet effect"." in [35].

[40] Strange, W. et al. (1993) "Consonant context affects perceived similarity of North German and American English vowels." *J. Acoustic. Soc. Am.*, 94, 1866.

WHAT DETERMINES THE PERCEPTUAL DIFFICULTY ENCOUNTERED IN THE ACQUISITION OF NONNATIVE CONTRASTS?

Ocke-Schwen Bohn
English Department, Kiel University, Germany

ABSTRACT

Of the several types of variables which interact to determine perceptual difficulty in the acquisition of nonnative contrasts, two types are selected for review: *Subject* variables, which define what the learner brings to the task of perceptually organizing nonnative contrasts, and *contrast* variables, which define what the learner is trying to organize perceptually.

INTRODUCTION

Probably anyone working in the field of L2 speech would agree that native language background is the one factor that contributes most importantly to perceptual difficulty in the acquisition of nonnative contrasts. Even though the evidence in support of this view is massive and unambiguous, such an answer would be simplistic, if not incorrect if provided without further qualification. This is so because L1 background is just one of many variables that may interact in complex ways to determine ease or difficulty in L2 speech perception. Polka [1] and Strange [2] have shown how the influence of this variable is mediated by, for instance, stimulus variables, which define *what* is selected for examination in L2 speech studies (e.g., type of contrast, acoustic cues to the contrast, phonetic contexts), and by task variables (including testing procedures), which define *how* the perception of nonnative contrasts is examined.

The organizational schema provided by Polka and Strange (orginally designed by Jenkins [3] to describe memory phenomena) will serve as a guideline to review what determines the perceptual difficulty encountered in the acquisition of nonnative contrasts. Because an overview of all variables and their interactions in L2 speech perception is beyond the scope of this contribution, it will focus on those types of variables that researchers *select* for study (i.e., subject and contrast variables), and largely ignore those variables that researchers can *manipulate* in studies of L2 speech perception (i.e., task variables). This focus is somewhat arbitrary and not intended as a comment on the relative importance of variables. The contributions of Strange and Jamieson emphasize the role of task variables (and their interactions with other variables) in L2 speech perception and in the training of nonnative contrasts. For detailed and comprehensive reviews of issues in L2 and cross-language speech perception, see [4].

SUBJECT VARIABLES

There are three subject variables whose roles in L2 speech perception have been studied and documented: L1 background (in great detail), L2 experience, and age of the learner (in somewhat lesser detail). Other variables must be involved, because large individual differences in the abilities of subjects to differentiate nonnative contrasts have frequently been observed even when subjects were homogeneous with respect to their L1 background, their L2 experience, and their age. However, the exact nature of these variables remains elusive. Some subject variables like gender, attitudes toward the L2 and the culture associated with the L2, or motivation to learn the L2, do not seem to contribute to the ability to differentiate nonnative contrasts. At present, the large individual differences especially among adult subjects can at best be attributed to some "talent" for language learning, but this cover term is clearly unsatisfactory because it is unknown what makes for a talented L2 perceiver. Large-scale correlational studies of L2 speech perception in adults are called for to identify which as yet unknown variables may play a role in creating the kind of outliers found in almost any L2 speech study, i.e., nonnative listeners with excellent perceptual abilities despite limited L2 experience at the one end and nonnative listeneners who seem to be immune to L2 experience at the other end.

L1 Background

Polivanov [5] was probably the first to acknowledge in detail the important role of L1 background in L2 speech perception. Polivanov does not clearly state the empirical bases for his observations; it appears that they are derived from nonnative speakers' productions (which are not a good indicator of perceptual problems, s. Llisterri, this volume). Still, Polivanov's insights anticipate the results of experimental work carried out almost 40 years later. His remarks on L1-dependent "thresholds of differentiation" (i.e., category boundaries) for initial stops differing in voice onset time (VOT) foresee one of the main results of Lisker & Abramson's [6] seminal study, in which listeners from three different language backgrounds (English, Spanish, and Thai) identified stimuli from a synthetic VOT-continuum in L1-specific ways: "the same pronunciation of the stop consonant ... will be relegated to different members of the given pair of phonemes in each of the ... given language consciousnesses". It is these different language consciousnesses, or, as one would put it today, language-specific ways to organize phonetic distinctions into phonologically relevant contrasts, that are a major source of perceptual difficulty in the acquisition of nonnative contrasts, as in Polivanov's example of Russian prevoiced /b/ and voiceless unaspirated /p/, which "seem completely identical for the Chinese perception".

A very large number of studies conducted over the past 25 years have documented that difficulty in the perception of nonnative contrasts is systematically related to the perceptual differentiation of phonetic contrasts in the L1. The influence of L1 background on L2 perception, however, is not all-pervasive. Depending on the nature of the variables that irteract with L1 background, its influence on the perception of nonnative contrasts may be reduced by specific L2 experience [7], it may be stronger for certain types of contrasts than for others [8], it can be amplified or attenuated through experimenters' choice among task variables [9], or it may even be completely absent for certain types of cues to the nonnative contrasts [10] and for nonnative contrasts that are nonassimilable to L1 contrasts [11].

L2 Experience

Studies which compared adult learners with varying amounts of L2 conversational experience for their ability to differentiate L2 contrasts indicate that L2 experience may induce L2 learners to reorganize their "linguistic consciousness". For instance, in a study that examined how two groups of L1 Japanese learners differing in English language experience identified and discriminated stimuli from a synthetic English /r/-/l/ continuum, MacKain et al. [7] reported that Japanese learners with extensive English experience had much steeper identification functions and higher and narrower discrimination peaks than L1 Japanese learners with little English language experience. However, the experienced Japanese learners of English still performed less accurately than native English listeners on discrimination of the English /r/-/l/ contrast (see also [12]).

In another case, Bohn & Flege [13] reported that L1 German learners of English, who had spent less than 1 year in the USA, differentiated synthetic stimuli from an English /ɛ/-/æ/ continuum by almost exclusively using temporal cues and ignoring spectral cues. Their perception of this L2 contrast was quite unlike a group of L1 English listeners, who relied almost exclusively on spectral cues to differentiate the /ɛ/-/æ/ contrast. Evidence of perceptual learning through L2 experience was provided by a group of German learners of English with extensive conversational L2 experience (>5 years of residence in the USA), who differentiated this contrast in a way that resembled native English listeners' perception in that they relied more on spectral than temporal cues.

However, as in other studies that examined the influence of L2 experience on the perception of L2 contrasts [7, 12], Bohn & Flege also found that several years of experience do not guarantee that L2 learners' perception will become completely native-like. The perception of

the /ɛ/-/æ/ contrast by the experienced German group was more English-like than that of the inexperienced German group, but it still differed from the native English listeners.

Why is it that even after years of experience, adult L2 learners' perception of L2 contrasts remains less accurate than the perception of native listeners? One factor that seems to contribute to perceptual difficulty despite massive L2 experience is the relation of L2 contrasts to native categories (s. also below). A number of studies carried out by Flege and his collaborators (summarized and reviewed in [14]) suggest that L2 experience is most likely to lead to perceptual learning if at least one of the members of the L2 contrast is "new", i.e., has no easily identifiable counterpart in the learner's L1, as English /æ/ for L1 German learners. If, however, both members of the L2 contrast are easily assimilable to counterparts in the L1 that are similar, but not identical to the members of the L2 contrasts, perceptual learning seems to be blocked.

This would explain, for instance, why experienced German learners of English in the Bohn & Flege study [13] did not differ from the inexperienced German learners in their use of both spectral and temporal cues to differentiate synthetic stimuli from an English /i/-/ɪ/ continuum. The reliance on both cues to differentiate the English /i/-/ɪ/ contrast seems to be a perceptual strategy transferred from differentiating German /i/-/ɪ/. Despite massive L2 experience, the experienced German learners did not even start to differentiate the English /i/-/ɪ/ contrast in a way that approximated the native English listeneners' perceptual strategy, which was characterized by the predominant use of spectral cues.

For easily assimilable L2 contrasts whose members have similar counterparts in the learners' L1, L2 experience in the form of conversational experience does not seem to induce perceptual learning. Still, training studies have yet to show whether this particular relation of L2 contrasts to native categories does indeed immunize L2 learners' perceptual learning abilities from structured L2 experience.

An important methodological problem in studying the role of L2 experience concerns the quantification of that variable, which is prerequisite to an adequate assessment of the quality and quantity of L2 input needed to induce perceptual learning. As a first approximation, length of residence in the L2 community would seem to be a valid measure of L2 experience. However, even learner groups with largely homogeneous social characteristics may differ considerably both in the amount of processible input they receive over a given time period and perhaps even in the quality of L2 input (e.g., authentic vs. foreign-accented). An obvious way to approach this problem is to collect detailed language background data from subjects, but how is one to weight different qualities and quantities of L2 input, e.g., amount of L2 mixed-dialect input at the workplace vs. foreign-accented input at home?

These methodological problems must be clarified before one can address the important question of what the limits of ultimate attainment are in adult L2 speech learning. For expample, Bohn & Flege [13] simply assumed that English language experience would be minimal for L2 learners who had spent an average six months in the USA, and that perceptual learning had reached its ultimate level for those L1 German learners who had spent an average 7 years in the USA. This assumption was justified to a large extent, for the two German groups differed clearly in how they perceived the English /ɛ/-/æ/ contrast. (They differed even more clearly in how they differentiated that contrast in production [15].)

Even though common sense would predict that the experienced learners in the Bohn & Flege [13] study had probably reached their level of ultimate attainment, L2 speech research still needs to address the questions of what amount of L2 experience enables perceptual learning and when this learning comes to a halt despite continued exposure to the L2. Studies of L2 learning in such diverse areas as speech production [16] and morphology and syntax [17] suggest that after a maximum of 5 years of L2 experience, adult L2 learners have reached their ultimate level of attainment.

One very interesting finding that has received relatively little attention

concerns the long-term effects of early experience. Tees & Werker [18] reported that subjects who had been exposed to Hindi in early childhood (but who had no contact with Hindi after age 2) could discriminate two Hindi stop contrasts (retroflex vs. dental and dental voiceless aspirated vs. voiced aspirated), whereas L1 English listeners with no Hindi experience performed very poorly. This suggests that early experience with a specific contrast helps maintain perceptual abilities necessary for discrimination of that contrast until much later in life even without intervening specific experience. In a study which examined the perception of a Salish place of articulation contrast (velar vs. uvular) in voiceless glottalized stops, Polka [19] reported that even though neither English nor Farsi has that specific Salish contrast, early Farsi bilinguals apparently benefited from nonspecific early experience with the Farsi velar vs. uvular contrast for voiced stops. This suggests that specific early experience may not be necessary to maintain accurate perception. Rather, broad experience with features employed to differentiate a contrast may be sufficient to maintain perceptual abilities.

Age

The most influential hypothesis on the age factor in language acquisition is Lenneberg's [20] Critical Period Hypothesis, which states that successful acquisition is possible only between the ages of (roughly) 2 and 12 years. This was hypothesized to be so primarily because only the prepubescent brain was supposed to have the plasticity needed to allocate new language functions. Even though research conducted over the past 25 years has shown that the original assumptions regarding the time frame, its biological basis, and the abruptness of the boundaries of the alleged critical period for language learning are wrong [16, 17, 21], the fact remains that children typically acquire languages with apparent ease and successfully, whereas language learning in adults is typically more effortful and, in the end, less successful.

These age-dependent differences are especially marked in speech perception. Research on infants' abilities to discriminate speech contrasts has shown that young infants (< six months of age) discriminate consonant contrasts in a categorical manner, no matter whether they had been exposed to the relevant contrasts in their ambient language [22]. The fact that young infants are not language-specific perceivers whereas adults are, lead to the assumption that adults' perceptual difficulties with nonnative contrasts were due to a loss of perceptual abilities with regard to those phonetic differences that are not phonologically distinctive in their L1.

Two types of evidence have shown that the inferior perceptual abilities of adults (as compared to infants) do not result from an atrophy of sensory abilities. First, Werker & Logan [23] showed that adults can discriminate acoustic differences that define nonnative contrasts if task variables are manipulated in a way that enables adult listeners to attend to stimuli in a general auditory rather than a specific phonetic mode. However, if adults process speech sounds in the specific phonetic mode of perception (which they normally do), they do not attend to acoustic detail that is irrelevant to category membership in the L1. It appears that L1 experience leads native listeners to focus on just those acoustic properties of speech sounds that define category membership in the L1. This selective attention is highly overlearned and indispensable for accurate and efficient perception of speech sounds in the L1, but it may entail, inattention to those acoustic dimensions and patterns that nonnative languages may employ to classify phonetic segments into functional categories.

Another type of evidence indicating that adult listeners' difficulties with nonnative contrasts are attentional in nature comes from studies which report successful learning for at least some nonnative listeners. For instance, Bohn & Flege [13] found that a sizeable proportion of experienced German learners had learned to differentiate the new English /ɛ/-/æ/ contrast in an English-like manner. No such evidence of perceptual learning was found for the similar English /i/-/ɪ/ contrast. The implication that learnability is a function of the relation of L2 contrasts to nonnative categories is incompatible with the view that perceptual problems of adults are due to sensory loss, for why

should this loss affect similar sounds of the L2, but not new ones?

CONTRAST VARIABLES

Studies which examined the perception of two or more contrasts by the same listeners using identical procedures typically report that nonnative contrasts differ both in the amount of difficulty they present initially and in their learnability [8, 11, 12, 13, 19, 24, 25]. What accounts for this nonuniformity? Is it the inherent salience of acoustic parameters that signal different types of contrasts, or is it L1 experience with certain contrasts (and features used to differentiate those contrasts) and the relation of L2 contrasts to L1 categories that determines relative difficulty? These questions will be addressed by looking at types of contrasts and types of cues for which differences in perceptual difficulty have or have not been reported.

Type of contrast

Different types of contrasts have received different amounts of attention in cross-language and L2 studies. Most studies have examined consonant contrasts, in particular the voicing contrasts in syllable-initial stops [6, 18, 22] and place-of-articulation contrasts [7, 8, 12, 19, 23]. With a few exceptions (e.g., [26]), nonnative vowel perception has only recently received detailed attention [10, 13, 25, 27, 28, 29].

Studies which compared nonnative perception of voicing and place-of-articulation contrasts have found that nonnative place contrasts are generally more difficult to differentiate and more resistant to learning than voicing contrasts. For instance, Tees & Werker [18] showed that a Hindi voicing contrast (voiceless aspirated vs. voiced aspirated) was easier to learn for L1 English subjects than a Hindi place contrast (retroflex vs. dental). One interpretation of this and of similar findings is that the voicing contrast is psychophysically more distinctive or robust than the complex spectral and temporal changes that signal place contrasts [30]. Alternatively, nonspecific L1 experience with the voicing contrast as opposed to lack of experience with place contrasts examined may account for these findings [31]. This may explain, for instance, why Jamieson & Morosan [32] found that L1 French listeners, whose L1 employs the voicing contrast in fricatives, learned to differentiate the English /θ/-/ð/ contrast rapidly, and why L1 Japanese learners have massive and persistant learning problems with English /r/-/l/ [7, 12, 31].

Nonnative perception of English /r/-/l/ also serves to illustrate a point that has only recently been studied in detail, namley, the phonetic and phonotactic context in which nonnative speech contrasts occur [8, 19]. For instance, studies by Pisoni and his collaborators (summarized in [33]) have shown how position (e.g., pre- vs. postvocalic) influences L1 Japanese listeners' ability to differentiate English /r/-/l/. In the postvocalic position, perception is much more accurate (because of the coloring of the preceding vowel) than in the prevocalic position. Phonetic context effects on nonnative vowel perception have recently been examined by Strange [34], who reported that the goodness of fit and the categorization of German /Y/ into English front or back vowel categories depended on the consonantal context in which /Y/ occurred. The results of the Strange [34] study suggest that reference to formant targets is not sufficient to explain patterns of interlingual identification for vowels.

In the discussion of the L1 background factor, it was mentioned that perceptual problems with consonant contrasts are largely predictable from the way in which the L1 classifies phonetic distinctions into functional categories. It is not clear whether this is also true of L2 vowel contrast perception. Rochet [28] reported that L1 speakers of English, of Portuguese, and of French labeled a high vowel continuum (/i/-/y/-/u/) in ways that directly reflected how their respective L1s use and segment that part of the vowel space. However, studies of vowel discrimination suggest that L2 vowel perception is less influenced by L1 background than L2 consonant perception. For instance, Stevens et al. [26] found that L1 background had little effect on how L1 Swedish and L1 English listeners discriminated isolated steady-state vowels. In addition, high discrimination levels for naturally produced nonnative vowels have been reported by Polka & Werker [27].

Clear differences in the ability to discriminate nonnative vowels have been observed in infant speech perception. In a cross-language study that examined the discrimination of the German-only contrast /u/-/y/ and the English-only contrast /ɛ/-/æ/ by English-learning and by German-learning infants in two age groups each (6-8 and 10-12 months), Polka & Bohn [35] found that the /u/-/y/ contrast was more discriminable for both language groups (and both age groups) than the /ɛ/-/æ/ contrast. For this case, at least, differences in the ability to perceive vowel contrasts seem to have a universal (e.g., psychophysical) rather than an experiential (i.e., L1 background) basis. Further research is underway to examine whether certain areas of the vowel space or certain acoustic dimensions that underlie vowel contrasts are more discriminable than others both in early infancy and adulthood [36].

Type of cue

Few studies have directly addressed the question of whether perceptual and learning problems are related to the nonnative-like use of cues that signal a contrast [10, 13, 37, 38, 39]. These studies typically employ the trading relations paradigm, in which redundant acoustic dimensions underlying a contrast are varied orthogonally in synthetic speech stimuli. For instance, Yamada & Tohkura [37] found that Japanese learners' perceptual problems with the English /r/-/l/ contrast are related to their use of the F2 transition cue, whereas native American English listeners predominantly use F3 onset frequency to differentiate /r/-/l/.

A set of studies examining trading relations in nonnative vowel perception was conducted by Flege and Bohn (summarized in part in [10]). L1 speakers of German, of Spanish, and of Mandarin who had limited L2 English experience were tested for their use of temporal vs. spectral cues in differentiating new English vowel contrasts (/ɛ/-/æ/ for L1 Germans, /i/-/I/ for L1 Spanish and Mandarin speakers). Native English listeners differentiated these contrasts almost exclusively on the basis of spectral differences, but the nonnative listeners responded primarily on the basis of duration rather than spectral differences. This perceptual strategy of inexperienced L2 listeners could not be attributed to the use of the duration cue in their respective L1s, for neither Mandarin nor Spanish differentiate vowels on the basis of duration. Bohn [10] hypothesized that the use of the duration cue to differentiate a new vowel contrast is an L1-independent, universal strategy that is applied whenever L1 experience has desensitized nonnative listeners to spectral differences in areas of the vowel space that are underexploited by the L1.

Further studies are needed to help determine which of the multiple cues signaling a nonnative contrast contribute to perceptual and learning problems, and what makes nonnative listener use cues that are not used by native listeners. An area of research that has only recently started to attract attention is the use and integration of visual cues in nonnative speech perception. In a study that examined cross-language influences on bimodal speech perception, Werker et al. [40] found an increasing relation between L2 English experience and the extent to which L1 French listeners integrated visual and acoustic cues. Training studies might profit from directing L2 learners' attention not just to critical acoustic cues, but also to visual cues which learners can exploit in face-to-face communication (e.g., visibility of the tongue tip in interdentals, which are acoustically very similar to labiodentals).

CONCLUSION

This review indicates that there is no simple answer to the question of what determines perceptual difficulty in the acquisition of nonnative contrasts. Two models have been proposed which pay tribute to the complex interactions of subject and contrast variables in L2 and cross-language perception. Both Best's [11, 29] Perceptual Assimilation Model (PAM) and Flege's [14] Speech Learning Model (SLM) attempt to predict perceptual difficulty on the basis of the perceived relation of nonnative speech sounds (i.e., contrast variables) to L1 categories (i.e., the L1 background variable). The models complement each other in that the SLM, which focuses on individual segments rather than contrasts, is a developmental model that

incorporates the subject variables L2 experience and age, whereas PAM is a model of cross-language perception that tries to account for listeners' initial difficulty with nonnative contrasts.

Flege's SLM classifies the relation between L1 and L2 sounds along a continuum ranging from "identical" over "similar" to "new". New sounds of the L2 are hypothesized to be sufficiently dissimilar from any L1 sound so that L2 learners will eventually discern the difference and establish a new perceptual category. Similar sounds, however, are classified by L2 learners as equivalent to their L1 counterparts, which blocks category formation.

According to Best's PAM, nonnative contrasts are assimilated to L1 categories either as good exemplars, acceptable exemplars, or notably deviant exemplars of the L1 category. In addition, nonnative categories that are very discrepant from any L1 sound are not assimilated into native categories at all, they are heard instead as some sort of nonspeech sound. Difficulty in the perception of nonnative contrasts is predicted by these different assimilation patterns.

Both models have been tested in several studies (reviewed in [14] and [29]) and the results have been, in general, quite supportive the models' predictions (but see [14]). One important problem which both Best and Flege acknowledge is that the predictive powers of both PAM and SLM rest upon the perceived phonetic similarity of L1 and L2 speech sounds. Progress in L2 speech perception research, which has come a long way since Polivanov [5], depends to a large extent upon success in developing objective means for predicting patterns of assimilation and interlingual identification.

REFERENCES

[1] Polka, L. (1989), *The role of experience in speech perception: Evidence from cross-language studies with adults*, Unpublished doctoral diss., University of South Florida, Tampa.

[2] Strange, W. (1992), "Learning non-native phoneme contrasts: Interactions among subject, stimulus, and task variables." Tohkura, Y. et al. eds., *Speech perception, production and linguistic structure*, Tokyo: Ohmsha, 196-219.

[3] Jenkins, J. J. (1979), "Four points to remember: A tetrahedral model of memory experiments". Cermak, L. S. & Craik, F. I. M., eds., *Levels of processing in human memory*, Hillsdale, NJ: Erlbaum, 429-446.

[4] Strange, W., ed. (1995), *Speech perception and linguistic experience: Theoretical and methodological issues*. Timonium, MD: York Press. (in press)

[5] Polivanov, E. D. (1931), "La perception des sons d'une langue étrangère." *Travaux du Cercle Linguistique de Prague* 4, 79-96. English translation in: Leont'ev, A. A., ed. (1974), *E. D. Polivanov: Selected works*, The Hague: Mouton, 223-237.

[6] Lisker, L. & Abramson, A. S., 1970, "The voicing dimension: Some experiments in comparative phonetics." *Proceedings 6th International Congress of Phonetic Sciences*, 563-567.

[7] MacKain, K., Best, C. & Strange, W. (1981), "Categorical perception of English /r/ and /l/ by Japanese bilinguals." *Applied Psycholinguistics* 2, 369-390.

[8] Polka, L. (1991), "Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions." *J. Acoustical Soc. America* 89, 2961-2977.

[9] Werker, J. F. & Tees, R. C. (1984), "Phonemic and phonetic factors in adult cross-language speech perception." *J. Acoustical Soc. America* 75, 1866-1878.

[10] Bohn, O.-S. (1995), "Cross-language speech perception in adults: L1 transfer doesn't tell it all." In [4].

[11] Best, C. T., McRoberts, G. W. & Sithole, N. N. (1988), "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants." *J. Experim. Psychology: Human Perception & Performance* 14, 345-360.

[12] Best. C. T. & Strange, W. (1992), "Effects of experience and phonetic factors on cross-language perception of approximants". *J. Phonetics* 20, 305-330.

[13] Bohn, O.-S. & Flege, J. E. (1990), "Interlingual identification and the role of foreign language experience in L2 vowel perception." *Applied Psycholinguistics* 11, 303-328.

[14] Flege, J. E. (1995) "Second language speech learning: Theory, findings and problems". In [4].

[15] Bohn, O.-S. & Flege, J. E. (1992), "The production of new and similar vowels by adult German learners of English." *Studies in Second Language Acquisition* 14, 131-158.

[16] Oyama, S. (1976), "A sensitive period for the acquisition of a nonnative phonological system." *J. Psycholinguistic Research* 5, 261-283.

[17] Johnson, J. S. & Newport, E. L. (1989), "Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language." *Cognitive Psychology* 21, 60-99.

[18] Tees, R. C. & Werker, J. F. (1984), "Perceptual flexibility: Maintenance or recovery of the ability to discriminate non-native speech sounds." *Canadian J. Psychology* 38, 579-590

[19] Polka, L. (1992), "Characterizing the influence of native language experience on adult speech perception." *Perception & Psychophysics* 52, 37-52.

[20] Lenneberg, E. H. (1967), *Biological Foundations of Language*, New York: Wiley.

[21] Flege, J. E. (1987), "A critical period for learning to pronounce foreign languages?" *Applied Linguistics* 8, 162-177.

[22] Lasky, R. E., Syrdal-Lasky, A. & Klein, R. E. (1975) "VOT discrimination by four to six and a half month old infants from Spanish environments." *J. Experim. Child Psychology* 20, 215-225.

[23] Werker, J, F. & Logan, J. S. (1985) "Cross-language evidence for three factors in speech perception." *Perception & Psychophysics* 37, 35-44.

[24] Werker, J. F. et al. (1981), "Developmental aspects of cross-language speech perception." *Child Development* 52, 349-355.

[25] Polka, L. (1995), Linguistic influences in adult perception of non-native vowel contrasts. *J. Acoustical Soc. America* 97, 1286-1296.

[26] Stevens, K. N. et al. (1969), "Crosslanguage study of vowel perception." *Language & Speech* 12, 1-23.

[27] Polka, L. & Werker, J. F. (1994), "Developmental changes in the perception of nonnative vowel contrasts". *J. Experim. Psychology: Human Perception & Performance* 20, 421-435.

[28] Rochet, B. L. (1995), "Perception and production of L2 speech sounds by adults." In [4].

[29] Best, C. T. (1995) "A direct realist view of cross-language speech perception". In [4].

[30] Burnham, D. K. (1986), "Developmental loss of speech perception: Exposure to and experience with a first language." *Applied Psycholinguistics* 7, 207-239.

[31] Strange, W. & Dittman, S. (1984), "Effects of discrimination training on the perception of /r-l/ by Japanese adults." *Perception & Psychophysics* 36, 131-145.

[32] Jamieson, D. G. & Morosan, D. E. (1986), "Training non-native speech contrasts in adults: Acquisition of the /ð/-/θ/ contrast by francophones." *Perception & Psychophysics* 40, 205-215.

[33] Pisoni, D. B. & Lively, S. E. (1995), "Methodological issues in training listeners to perceive nonnative contrasts." In [4].

[34] Strange, W. et al. (1993) "Consonant context affects perceived similarity of North German and American English vowels." *J. Acoustical Soc. America* 94, 1866.

[35] Polka, L. & Bohn, O.-S., (in prep.) A cross-language comparison of vowel perception in infancy.

[36] Bohn, O.-S. & Polka, L. (1995), "What defines vowel identity in prelingual infants?" *Proceedings 13th International Congress of Phonetics.*

[37] Yamada, R. & Tohkura, Y. (1992), "Perception and production of syllable-initial English /r/ and /l/ by native speakers of Japanese. *Proceedings International Conference on Spoken Language Processing*, 757-760.

[38] Gottfried, T. L. & Beddor, P.S. (1988), "Perception of temporal and spectral information in French vowels." *Language & Speech* 31, 57-75.

[39] Hazan, V. L. & Boulakia, G. (1993), "Perception and production of a voicing contrast by French-English bilinguals". *Language & Speech* 36, 17-38.

[40] Werker, J. F., Frost, P. E. & McGurk, H. (1992), "La langue et les lèvres: Cross-language influences on bimodal speech percpetion." *Canadian J. Psychology* 46, 551-568.

RELATIONSHIPS BETWEEN SPEECH PRODUCTION AND SPEECH PERCEPTION IN A SECOND LANGUAGE

Joaquim Llisterri

Departament de Filologia Espanyola, Facultat de Filosofia i Lletres, Edifici B, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.
Fax: +34.3.581.16.86; E-mail: joaquim.llisterri@cc.uab.es

ABSTRACT

This contribution presents a review of some of the findings reported in the literature concerning the relationship between production and perception of the sounds of a second language. Part of the experimental research and some applied work is summarized. It is concluded that a complex relationship exists between the production and perception of L2 sounds, and that many factors have to be taken into account in describing this interaction.

INTRODUCTION

Research concerned with the relationship between the production and perception of the sounds of a second language (L2) has addressed a problem that can be summarized as follows: Does production precede perception or, conversely, does perception precede production in the process of acquiring an L2? That is to say: can learners adequately pronounce sounds which are not well perceived, or is a good perception a prerequisite to accurate pronunciations?

The answer to this question has not only got theoretical implications regarding the process of L2 acquisition, but also practical consequences as far as the methodology used for teaching pronunciation is concerned.

Perception precedes production

As early as in 1931, Polivanov [1] claimed that the phonemic representations of a second language are perceived according to the system of the first language; although it is difficult to assess the validity of the data supplied by Polivanov (see Rochet [2] note 3), his remarks have been interpreted as supporting the hypothesis that difficulties in the production of the sounds of an L2 arise from the influence of the L1 phonological structure on the perception of L2 sounds.

A very similar view has been put forward by Trubetzkoy [3], who conceived the phonological system of L1 as a 'filter' through which all the sounds of L2 are perceived and classified. The verbo-tonal system closely follows this approach and, consequently, the principle orienting its methodology is that L2 sounds are not adequately produced because they are not correctly perceived [4]

Later on, the idea that inaccurate perceptual representations are responsible for non-native productions has been formulated in many of Flege's contributions. It can be summarized as follows: "foreign accent [...] may instead result from the development of the L1 phonetic system, which makes it increasingly unlikely that similar sounds in an L2 will evade being equated with sounds in L1" [5] (p. 285). This phenomenon has been defined as "equivalence classification".

Then, according to the hypothesis of the 'phonological filter' and the 'equivalence classification' principle, perception of a new phonetic contrast must necessarily precede its production.

Production precedes perception

As Borrell [6] points out, it is a very common experience when learning an L2 that not all the sounds that are correctly perceived will be correctly produced. Similar observations have been made by Neufeld [7] and by Brière [8]. It seems then, that in certain cases, the production of L2 sounds might precede their perception.

In order to find evidence supporting one or both principles outlined below, we will review part of the literature concerned with the relationship between speech perception and speech production both in L2 and in bilinguals.

The review will be divided in two parts: in the first one, the major findings of selected experimental studies concerning segmental elements will be summarized. The second part will concentrate on work carried out from a more pedagogical perspective, trying to establish the efficiency of pronunciation teaching methods that emphasize initial training either in production or in perception.

Studies addressing this topic in the first language will not be mentioned, but it is worth noting that "it is not clear how production and perception are linked in mature native speakers" [9] (p.51). The literature on L1 development will not be surveyed either, although undoubtedly it may shed some light into the issue.

Another field that should be taken into account but will not be discussed here is speech pathology. It should be reminded that the proponents of the verbo-tonal method have advocated that the situation of a non-native with regard to a foreign language is similar to the situation of a deaf person, in the sense that lack of an adequate perception inhibits accurate productions. Thus, the notion of 'phonological deafness' has been applied to second language learners in this context [5].

EXPERIMENTAL EVIDENCE OF RELATIONSHIPS BETWEEN PRODUCTION AND PERCEPTION IN L2

The results that can be found in the literature will be presented separately for vowels and for consonants since, as will be shown later, it has been proposed that different classes of sounds may behave in a different way concerning the relationship between production and perception [9].

The production and perception of L2 vowels

The experimental studies concerning vowel production and perception have consistently shown a close link between those abilities in L2 learners.

The strength of this link has been emphasized, for example, by Barry [10]. After examining the correlation between production and perception of English vowels by German speakers, he concluded that "well-established perceptual categories are more likely to be accompanied by more acceptable production" (p.160), suggesting the possibility of using perceptual abilities to predict accuracy in production.

The perceptual bases of production errors have been also examined in a recent experimental study by Rochet [2]. Speakers of Canadian English and of Brazilian Portuguese were asked to imitate the French vowel [y] and to label a continuum of synthetic high vowels as [i] or [u]. Production errors - systematic substitutions of French [y] by [u] in the case of English speakers, and by [i] in the case of Portuguese speakers - were correlated to labelling results in a perceptual test: as predicted by the hypothesis, vowels with a second formant in the range of French [y] were mostly identified as [u] by English speakers and as [i] by Portuguese speakers.

According to Rochet, these results support the notion that "accented pronunciations of L2 sounds by untrained speakers may be perceptually motivated"; they also seem to imply a good correlation between perception and production, the first taking precedence over the second.

However, other studies suggest that the perception-production link is of a more complex nature, and might be influenced by different factors.

Elsendoorn [11] compared the results of an auditory task in which Dutch learners of English had to adjust the duration of vowels in English words with durational data obtained from words produced by the same group of subjects. The results showed an effect of the fortis or lenis character of the final obstruent in the perceptual task that was not present in the production data.

Another interesting result of the experiment was that standard deviations of perceptual adjustments diminished "with growing familiarity with and knowledge of the English language" (p.675). This introduces a new dimension that increases the complexity of the production-perception link, namely the knowledge of L2.

Bohn & Flege [9] directly approached the influence of this factor by examining the production and perception of the English vowels /ɛ/ and /æ/ in two groups of German learners of English: an experienced group - consisting of

individuals who had lived in the United States for at least five years - and an inexperienced one - formed by learners with a mean length of stay in the US of six months -.

The results of the experiment revealed a clear difference between the two groups of speakers. The inexperienced German learners did not produce the contrast between the two vowels under consideration, but they succeeded in differentiating the vowels in a labelling task; however, they relied primarily in duration, a cue that does not strongly influence the labelling of the same continuum by native English speakers. Experienced German speakers of English did produce the contrast between the two vowels and were able to achieve better labelling results than the non-experienced ones, but they did not use durational and spectral cues in the same way that native English speakers did in the same labelling task.

Bohn & Flege also remarked that while spectral differences in production were related to a dominance of spectral cues in perception, strong reliance on durational cues in perception implied small durational differences in production.

These results point out that more elements are to be taken into account when attempting to explain the relationship between production and perception in vowels: on the one hand, experience with the language seems to have a more marked influence on production than on perception, confirming Elsendoorn's findings [11]; on the other hand, the behaviour of the speakers examined shows that "different relations between production and perception may exist for different acoustic correlates (such as vowel spectrum and duration) of phonetic categories" [9] (p.52). Then, not only contextual dependency may induce different behaviours in L2 production and perception, but also different acoustic cues seem to be used in different ways depending on the activity that is carried out by an L2 learner.

As far as the precedence of perception over production is concerned, Bohn & Flege concluded from their experiment that "in the early stages of L2 speech learning, perception may [...] lead production, although the perceptual criteria may be very different from the ones used by native speakers" (p. 52).

However, the fact that the inexperienced German speakers of English were able to differentiate perceptually vowels that they were not able to contrast in production [9], and the fact that both the Portuguese and the English speakers studied by Rochet [2] correctly imitated the French vowel [y] in roughly 50% of their French productions seems to imply that production abilities can not be completely inferred from perceptual ones and vice-versa. In fact, Bohn & Flege [9] found that experience might improve production, since "continued L2 contact enables L2 speakers to produce a new vowel contrast like native speakers of the L2"; but at the same time, they noted that "perception abilities for a new vowel contrast may lag behind even after several years of L2 experience" (p.52)

Moreover, even if discrimination was correct, Bohn & Flege convincingly argued that the strategies used by L2 learners (experienced or not) might not be the same as the ones used by native speakers. (See also Bohn, this volume, for an explanation of the results)

In summary, as far as vowels are concerned, there is a complex link between production and perception in L2 sounds. Although it seems that perception in general might precede production, direct inferences about pronunciation accuracy can not probably be made from perceptual abilities in a straightforward manner. Factors such as contextual dependency, nature of the acoustic cues involved in phonetic contrasts and the learner's familiarity with the L2 must as well be taken into account.

The production and perception of L2 consonants

In order to exemplify the relationship between production and perception of consonants in a second language, we will summarize some findings concerning two widely studied classes: liquids and stops.

The studies concerning the production and perception of liquids have concentrated on the distinction between English /r/ and /l/ by speakers of languages such as Japanese and Korean, which do not contrast these segments in their phonological system.

An experiment was carried out by Sheldon & Strange [12] with Japanese speakers of English living in the United States. It was shown that the production of the English contrast between /r/ and /l/ was more accurate than the perception of natural utterances; the materials of the perceptual test comprised minimal pairs with /r/ and /l/ taken both from native productions and also from the subjects' own productions of the pairs.

The same findings were reported by Goto [13] in a previous experiment carried out in Japan; as concluded by Sheldon & Strange "at least for the contrast studied here, perceptual mastery of a foreign contrast does not necessarily precede adult learners' ability to produce acceptable tokens of the contrasting phonemes, and may, in fact, lag behind production mastery" (p.245).

Flege [5] discusses some of the factors that may have influenced the results reported by Sheldon & Strange: explicit articulatory training undergone by some of the speakers (see also the results of Catford & Pisoni's [14] experiment reported below), characteristics of the corpus - i.e., words read in citation form - and the monitored nature of the situation. Sheldon & Strange themselves noted that the group of speakers had very specific characteristics, in the sense that they were at an advanced state of L2 acquisition.

Borden, Gerber & Milsark [15] examined the relationship between perception and production of English /l/ and /r/ in Korean learners of English in an experiment that comprised production, identification, discrimination and a self-perception test. Synthesized stimuli were used for the identification and discrimination tests. One of the main results obtained was that self-perception develops earlier and may be a prerequisite for accurate production.

The authors note that "the ability to make phonemic perceptual judgments in an /r/ - /l/ continuum that are similar to those of English speakers also seems to improve before production" (p. 516). This results for Korean learners seem to be at least in partial disagreement with those reported by Sheldon & Strange [12] for Japanese speakers.

However, the results obtained by Borden *et al.* have been reanalyzed by Sheldon [16]. The statistical treatment that Sheldon applied to the data allows for an interpretation of the results that is more coherent with the findings of Sheldon & Strange [12], since the idea that self-perception precedes production is not confirmed.

One of the important conclusions of Sheldon's reanalysis was that the relationship between production and perception depended on the amount of time spent in the USA by the Korean learners, so that "as the learner's time in the US increases the probability of occurrence of perception exceeding production decreases" (p.111). This would bee in agreement with the fact that the speakers studied by Sheldon & Strange [12] were advanced learners.

The model proposed by Sheldon is not only corroborated by her interpretation of the Borden *et al.* data, but also by the fact that, as she hypothesizes, a functional perceptual level in an L2 might be enough for communication purposes, while heavily accented productions are socially less accepted, with the consequence that L2 speakers would have more pressure to improve production than perception. The same conclusions from an experiment concerned with vowels are put forward by Bohn & Flege [9]: "Perhaps perception of a new vowel contrast is more resistant to L2 experience than production because speech production is more subject to social control than speech perception" (p.52).

Familiarity with the language appears then to be a factor that, as we have seen for vowels, may heavily determine the relationship between production and perception and may contribute to changes in these abilities over time.

We may now examine some of the results reported for the production and perception of stop consonants in L2.

Flege & Eefting [17] studied the production and perception of the /t/ - /d/ contrast by Dutch learners of English with different degrees of familiarity with the language. The results showed that

Dutch speakers were able to produce a substantial VOT difference between Dutch and English, indicating a good discrimination of the two languages. However, in perception tests, even the most proficient Dutch speakers showed only a small shift in the location of phoneme boundaries when identifying stops in a /dɑ/ - /tɑ/ synthetic continuum, which they were induced to perceive as Dutch or English by modifying the language setting of the experiment. This seems to suggest that the distinction between the two languages in perception was not as clear as in production. The discrepancies between the cross-language shift in production and perception lead the authors to conclude that there is a disparity between production and perception.

The same disparity can be observed in a study that directly addressed the influence on production of adequate perceptual representations of the sounds of L2. Flege [18] examined vowel duration as a cue to voicing in English words ending with /t/ or /d/ produced and perceived by Chinese speakers of English; subjects were classified according to age of L2 acquisition and experience with the language.

The existence of a link between production and perception was revealed by the correlations between differences in perceived vowel duration and degree of foreign accent judged by native speakers of English; correlations between voicing effects in production and differences in vowel duration in perception were found too. Nevertheless, it appeared that "non-natives will resemble native speakers more closely in perceiving than in producing vowel duration differences" [18] (p.1605), a result that would be in agreement with the 'perception before production' hypothesis, specially in the case of experienced learners.

However, it can be concluded from the study of individual differences that, in adult learners, production is not limited or inhibited by the perceptual representations of the L2 sounds; the explanation proposed by Flege to account for small differences in vowel duration found in some of the subjects together with large effects in perception are based on problems with the timing of articulatory gestures and on a reduced sensitivity to vowel duration differences due to self-hearing (since a correlation was found between unsuccessful imitations and strong foreign accent).

The conclusions of the Borden *et al.* study [15] about the importance of self-perception in the development of production accuracy and the possibility of developing a near-native perceptual ability before reaching the same level in production seem to agree with some of Flege's findings.

Finally, it is hypothesized by Flege, that the presence of more experienced learners in the group of subjects would have lent support to the precedence of perception over production.

Another study on the perception and production of Italian stop consonants by Austrian German learners carried out by Grasseger [18] also supported the hypothesis that well-established perceptual categories do help accurate productions. He also suggested, as Barry [10] did for vowels, that perceptual tests might be a good tool to indicate production difficulties.

It can be seen again that, as far as consonants are concerned, it is not easy to establish a direct correlation between production and perception in an L2 although obviously some links do exist.

The nature of this link does not seem clear if we try to integrate the experimental results. On the one hand, Sheldon & Strange [12] showed that production can precede perception in advanced learners; the Dutch speakers of English examined by Flege & Eefting [17] presented a better differentiation of the two languages in production than in perception; also, some of the late learners examined by Flege [18] showed larger production effects than perception effects. Sheldon's [16] hypothesis about the greater social pressure to improve production than perception could be a plausible explanation for these facts.

On the other hand, the findings of Borden *et al.* [15] seem to point out in the direction of a precedence of perception over production and the group results obtained by Flege [18] also seem to suggest a more native-like behaviour in perception than in production.

Finally, it should be noted that, as in the case of vowels, the age of L2 acquisition, the degree of exposure to the language, and the experience with L2 seem to be factors that affect the general correlation between production and perception.

Detection of foreign accent

Another useful contribution to the debate on the precedence of perception over production can be found in studies concerning foreign accent detection. Early work by Flege [20] will be summarized as an example.

Comparison of the performance of a group a Taiwanese subjects who had lived in the US for one year and another group who had lived there for five years showed that, in a foreign accent detection task, the experienced group was able to distinguish non-native from native speakers of English better than then non-experience group. However, both groups were rated as having equally strong foreign accent by native English judges.

According to Flege, this shows a dichotomy between speech production and speech perception, and it can be convincingly argued that, for the subjects of the experiment, their ability to detect non-authentic productions was greater than their ability in production.

Flege, then, proposed that perception is more subject to improvement with time than production is. However, the conclusions of Bohn & Flege's study of vowels [9] suggesting that "perception abilities for a new vowel contrast may lag behind even after several years of L2 experience" (p.52) do not seem to agree with the earlier findings; Sheldon's explanation [16] based on the social need to improve production more than perception seems to be more coherent with the 'production precedes perception' hypothesis.

Production and perception of L2 sounds in bilingual speakers

To conclude this part of the review, some experimental studies concerning the relationship between production and perception in bilingual speakers will be presented.

Caramazza *et al.* [21] compared the voiced/unvoiced contrast in stops produced and perceived by Canadian French-English bilinguals with the results obtained from monolingual speakers. They found that "the bilingual subjects produced voicing distinctions which were clearly different for the two languages; and this disparity stands in marked contrast with the similarity of their perceptual functions in the two language modes" (p.425); their conclusion was that bilingual speakers can better adapt their production than their perception in their non-dominant language. (See Hazan & Boulakia [22] for the effect of language dominance on bilingual's performance).

Similarly, in a study of the production and perception of English /t/ -/d/ and /i/ - /ɪ/ contrasts in early French-English bilinguals with English dominance, Mack [23] found evidence that "bilingual production can be more accurate than perception" (p.197).

The findings of both studies seem to point out towards a better differentiation between the two languages in production than in perception. The explanation suggested by Mack is similar to the one proposed by Sheldon [16] to account for the same trend in L2 speakers: the social consequences of non-native production are more important than those of non-native perception and, therefore, accurate productions are found whereas perception can be different from monolinguals, whenever comprehension is achieved.

PRODUCTION, PERCEPTION AND PRONUNCIATION TEACHING

Having reviewed some of the evidence found in the experimental literature, it might be useful to consider studies that have approached our topic from a perspective more oriented towards the teaching of pronunciation. The techniques for training non-native contrasts will not be presented here (see Jamieson, this volume); instead, we will concentrate on work discussing the effects of training based on production *vs.* training based on perceptual strategies.

In a classical experiment, Catford & Pisoni [14] showed a superiority in production and perception of a set of so-called 'exotic sounds' in subjects who received an articulatory training compared to subjects that were trained

with auditory techniques based on perceptual discrimination.

This led the authors to conclude that "what is effective in the teaching of sound production is the systematic development by small steps from known articulatory postures and movements to new and unknown ones" (p.477), implying, at the same time, that good production abilities may contribute to a better discrimination of L2 sounds.

The same conclusion seems to be supported by Weiss [24], reporting a previous experiment in which it was shown that training in pronunciation improved the discrimination abilities of a group of Chinese students of English. Another interesting result of this study was that a greater experience in a second language improves perception more than production (cf. [20] discussed earlier for a similar result)

It seems, then, that it can be provisionally concluded that training in production might help to improve perception. However, according to the results provided by Rochet [2], the opposite also seems to hold true.

Rochet claimed that a significant improvement in the production of French voiceless stops by native speakers of Mandarin Chinese - from 30% to 19% of incorrect pronunciations - was found after perceptual training with synthetic stimuli. This suggests that "improvement in perception performance can in turn translate into improvement in production performance".

The results from the experiment on the perception and production of /r/ and /l/ in English by Japanese speakers performed by Sheldon & Strange [12] should also be reminded in this context, since the authors argued that, at least for some consonants, good perception does not necessarily imply an accurate production and that "drills in production do not necessarily benefit (auditory) perceptual learning" (p.257).

This very brief review shows again that perception and production abilities in L2 are closely linked, but precedence of one over the other in training is not clearly established.

CONCLUSIONS

This necessarily non exhaustive review of some of the studies that have addressed the topic of the relationship between production and perception of the sounds of a second language has tried to show the complexity of the topic. However, some general trends can be signaled:

- It does not seem possible to infer production abilities from perceptual ones and vice-versa.
- Stage in the acquisition of L2, experience with the language, degree of exposure, and age of acquisition seem to play a major role in the interaction between production and perception in L2.
- The relation between production and perception might differ according to the class of sounds, to the acoustic and perceptual correlates of these classes and to contextual effects.
- Similarity between L1 and L2 sounds might also have an effect on the interplay between production and perception.
- Social factors such as pressure to improve production may provide an explanation for cases in which production precedes perception.

Moreover, methodological problems have to be considered. First of all, Mack [23] has mentioned the inadequacy of comparing results from tests in speech production with results derived from speech perception tests, since there are important differences in the nature of the techniques used to assess these activities. Secondly, Bohn & Flege [9] point out the difficulties in defining the criteria used to evaluate accuracy in production and perception, which lead to difficulties in comparing different studies. Finally, it has to be reminded that most of the work carried out has concerned experimental situations in which highly controlled tasks and linguistic materials have been used.

It seems, in summary, that we have come a long way in the characterization of production and perception skills, but as a recent paper by Rochet [2] concludes, "The relationship between the perception of L2 speech sounds and their production by non-native speakers is still far from well understood".

ACKNOWLEDGMENT

Thanks are due to Ninon Font and Astrid Roig for their help in the survey on which this paper is based.

REFERENCES

[1] POLIVANOV, E. (1931) "La perception des sons d'une langue étrangère" *Travaux du Cercle Linguistique de Prague* 4; in *Le Cercle de Prague* (Change, 3) Paris, 1969. pp. 111-14.

[2] ROCHET, B.L. (in press) "Perception and production of L2 speech sounds by adults" in STRANGE, W. (Ed) *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research*. Timonium, MD: York Press Inc.

[3] TRUBETZKOY, N.S. (1939) "Grundzüge der Phonologie", *Travaux du Cercle Linguistique de Prague*, 7; French translation by J. Cantineau: *Principes de phonologie*. Paris: Klincksieck, 1949.

[4] RENARD, R. (1979) *Introduction à la méthode verbo-tonale de correction phonétique*. Troisième édition entièrement refondue. Bruxelles: Didier.

[5] FLEGE, J. E. (1991) " Perception and Production: The Relevance of Phonetic Input to L2 Phonological Learning" in HUEBER, T.- FERGUSON, C. (Eds.) *Crosscurrents in Second Language Acquisition and Linguistic Theories*. Amsterdam: John Benjamins. pp. 249-289.

[6] BORRELL, A. (1990) " Perception et (re)production dans l'apprentissage des langues étrangères. Quelques réflexions sur les aspects phonético-phonologiques", *Revue de Phonétique Appliquée* 95-96-97:107-114.

[7] NEUFELD, G.G. (1988) "Phonological asymmetry in second language learning and performance", *Language Learning* 38,4: 531-559

[8] BRIÈRE, E. (1966) "An Investigation of Phonological Interference", *Language* 42,4: 769-796

[9] BOHN, O.S.- FLEGE, J.E. (1990) "Perception and Production of a New Vowel Category by Adult Second Language Learners" in LEATHER, J.- JAMES, A. (Eds.) *New Sounds 90: Proceedings of the 1990 Amsterdam Symposium on the Acquisition of Second-Language Speech*. Amsterdam: University of Amsterdam Press. pp. 37-56.

[10] BARRY, W. (1989) "Perception and production of English vowels by German learners : instrumental - phonetic support in language teaching", *Phonetica* 46:155-168

[11] ELSENDOORN, B.A.G. (1984) "Production and Perception of English Vowel Duration by Dutch Speakers of English" in VAN DEN BROECKE, M.P.R - COHEN, A. (Eds.) *Proceedings of the Tenth International Congress of Phonetic Sciences*. Dordrecht: Foris. pp. 673-676.

[12] SHELDON, A.- STRANGE, W. (1982) "The acquisition of /r/ and /l/ by Japanese learners of English: evidence that speech production can precede speech perception", *Applied Psycholinguistics* 3: 243-261

[13] GOTO, H. (1971) "Auditory perception by normal Japanese adults of the sounds "l" and "r" ", *Neuropsychologia* 9: 317-323

[14] CATFORD, J.C.- PISONI, D. (1970) "Auditory vs. Articulatory Training in Exotic Sounds", *Modern Language Journal* 54: 477-481

[15] BORDEN, G.- GERBER, A.- MILSARK, G. (1983) "Production and Perception of the /r/-/l/ Contrast in Korean Adults Learning English", *Language Learning* 33, 3: 499-526.

[16] SHELDON, A (1985) "The relationship between production and perception of the /r/-/l/ contrast in Korean Adults learning English. A reply to Borden, Gerber and Milsark", *Language Learning* 35, 1: 107-113.

[17] FLEGE, J.E. - EEFTING, W. (1987) "Cross-language switching in stop consonant perception and production by Dutch speakers of English", *Speech Communication* 6,3: 185-202

[18] FLEGE, J.E. (1993) "Production and perception of a novel, second-language phonetic contrast", *Journal of the Acoustical Society of America* 93,3: 1589-1608

[19] GRASSEGER, H. (1991) "Perception and production of Italian plosives by Italian learners" in *Actes du XIIème Congrès International des Sciences Phonétiques*. Aix-en-Provence: Université de Provence, Service des Publications. vol. 5, pp. 290-293

[20] FLEGE, J.E. (1988) "Factors affecting degree of perceived foreign accent in English sentences", *Journal of the Acoustical Society of America* 84, 1: 70-79

[21] CARAMAZZA, A.- YENI-KOMSHIAN, G.- ZURIFF, E.- CARBONE, E. (1973) "The Acquisition of a New Phonological Contrast: The Case of Stop Consonants in French-English Bilinguals", *Journal of the Acoustical Society of America* 54, 2: 421-428.

[22] HAZAN, V.- BOULAKIA, G. (1993) "Perception and production of a voicing contrast by French-English bilinguals", *Language and Speech*, 36,1: 17-38.

[23] MACK, M. (1989) " Consonant and vowel perception and production: Early English-French bilinguals and English monolinguals", *Perception & Psychophysics*, 46,2: 187-200.

[24] WEISS, W. (1992) "Perception and Production in Accent Training", *Revue de Phonétique Appliquée*, 102: 69-81.

TECHNIQUES FOR TRAINING DIFFICULT NON-NATIVE SPEECH CONTRASTS

D.G. Jamieson
Hearing Health Care Research Unit
Elborn College, University of Western Ontario
London, ON CANADA, N6G 1H1

Abstract
This paper examines the prospects for applying systematic procedures to train adult second language (L2) learners to perceive new speech contrasts. Four techniques that can produce generalized improvements in such speech perception are reviewed. Future research to optimize such training is then discussed.

Introduction
For adult second language (L2) learners, speech perception and production are strongly influenced by one's first language (L1). Early reports that training did not improve performance substantially, were seen as evidence that the capacity to learn new speech contrasts declined irreversibly, with age.

Within the past decade, new studies applying systematic approaches to train non-native speech contrasts have demonstrated that there is considerable ability to learn to perceive new speech contrasts, at least until early adulthood [1,2]. While the body of this work remains quite limited, the progress already made is extremely encouraging. This paper summarizes four successful training approaches that can facilitate the acquisition of L2 speech contrasts, and attempts to identify promising directions for future work in this field.

Effects of L1 on L2 Performance
When substantial exposure to an L2 is delayed until adulthood, one's ability to perceive and produce certain L2 speech sounds will be limited. These effects involve complex interactions among the learner's age when substantial exposure to the L2 begins, the sound patterns of the L1 and L2, the amount and type of exposure to L2, and the listener's individual perceptual skills and learning abilities [3,4,5]. For example English "th" voicing distinctions cause difficulty for native speakers of French; English /r/ and /l/ are difficult for native speakers of Japanese, Korean, and Cantonese; and the Hindi dental-retroflex consonant and French [u]-[y] vowel distinctions are difficult for English speakers [1,6,7,8].

Such difficulties may remain even after many years in the L2 environment, and they appear to be little influenced by traditional language training. When measured in terms of the accuracy with which the target sounds can be identified in high-quality recordings of isolated L2 words, under favourable listening conditions, error rates of 20% to 40% or more are common even for adults with several years of L2 experience; native speakers achieve virtually 100% accuracy under these circumstances. Moreover, the performance of L2 learners declines rapidly in more typical real-life listening environments, such as in multi-talker noise or reverberation [9].

More detailed consideration of the complex interactions among L1, subject, experience, and L2 variables is provided in the other papers from this session. Relevant data are also provided in [3,10,11] and useful theoretical perspectives are provided in [7,12].

Development of Speech Perception Abilities
Developmental studies of speech perception abilities have established the general principals that: 1) very young infants can discriminate most phonetic contrasts; 2) the ability to discriminate non-native contrasts declines rapidly within the first year of life; and 3) reduced discrimination ability reflects a change in attention to acoustic cues, rather than a loss of sensitivity to the acoustic cues [13]. Developmental studies therefore encourage the notion that many difficulties with L2 speech perception may be overcome with appropriate training techniques.

Characteristics of Successful Training Approaches
Much work remains to be done to optimize training techniques for new speech contrasts. However, we do know that training is more likely to be effective when the training task is designed to:

1) **focus the listener's attention on how acoustic patterns are mapped into phonemic categories in the L2.** That is, training tasks should direct attention to the acoustic factors that define each L2 category, while suppressing attention to acoustic cues that are irrelevant for classification in L2 [1,2]. This focus can be achieved by requiring listeners to identify (categorize) target sounds from among a set of candidate tokens. Discrimination tasks will not normally improve listeners' abilities to categorize sounds using L2 categories, as such tasks encourage attention to even phonetically-irrelevant differences between stimuli.

2) **provide prompt, unambiguous feedback concerning the L2 category appropriate to each training token.** Effective feedback can simply inform the listener about the accuracy of each judgement as soon as it is made. For example, the correct response can be indicated using a light, or by otherwise highlighting the display. One or more repetitions of the correct signal may also accompany this indicator. Mere listening to a sequence of samples from a specified category may also be effective [14].

3) **expose subjects to an adequate range of acoustic variation during training.** Subjects must learn not only about the acoustic cues that define and differentiate categories in the L2, but also about the range of variation that is tolerated within each category [1,15]. If enough is known about the relevant acoustic cues, synthetic speech signals may be used to direct listener attention to certain important cues from the outset of training. Alternatively one may use multiple tokens of the target sounds, spoken by several talkers, to provide a range of variation in the training set. Failure to provide sufficient variation restricts learning and/or reduces transfer outside the training environment [15, 16].

Measuring Performance of L2 Learners
A fundamental consideration for studies of L2 acquisition is how the L2 learner's speech perception and/or production performance will be measured. Assessment of performance typically focuses on the subject's ability to identify the presence of sounds from each of a few target categories, when listening to isolated L2 words. Production is assessed more rarely, and typically involves rating the quality of the learners' utterances, by native speakers.

Training might certainly be expected to improve performance when the test task and conditions are the same as those used in training. It is therefore important to ask **to what extent does training generalize to new and different tasks and conditions?** For example, to what extent does learning generalize to new speakers, to new words containing the target sounds, to words in which the target sounds occur in new (untrained) phonetic environments, to target sounds that share a feature with the training sounds, to other listening conditions (e.g., in conversation, in fluent speech, or when listening to degraded speech), and to the production of target sounds under various conditions? It is also important to establish the extent to which performance changes endure once formal training has been discontinued.

This review considers training techniques that have demonstrated such generalization of learning.

Successful Training Techniques

A fundamental finding of cross-language speech research is that many production difficulties have an underlying perceptual difficulty. The major effort has therefore been directed to improving the learner's speech perception abilities. At least four training techniques are known to produce relatively rapid improvements in listeners' abilities to perceive non-native speech contrasts. Studies have included several L1 groups and L2 targets.

Subject and specific language variables aside, the approaches can be differentiated in terms of several important variables: 1) the training task (eg., identification training); 2) the type and sequencing of training signals (eg., tokens from multiple talkers); and 3) the form of informational feedback used. These approaches are discussed in turn, below. Table 1 summarizes results from studies with these successful approaches.

Table 1
Summary of Training Studies Demonstrating Significant Improvement with Transfer to New Speakers

	L1 Group	L2 Target	Task	# Sess	Total Time (h)	# Trials	# Wks	Init. Perf. (%)	Final Perf. (%)	(%) Chg.	New Spk (%)
J&M '86	Fr.	/θ/-/ð/	F	2	~1.5	~720	1	62	92	48	-14
M&J '86	Fr.	/θ/-/ð/	F	2	~1.5	~720	1	61	92	52	-11
J&M '92	Fr.	/θ/-/ð/	F	4	<4	~720	1	70	89	27	-8
J & Y '95	Kor.-y	E /r/-/l/	NT	15	~4	1500	3	69	90	32	3
J & Y '95	Kor.-o	E /r/-/l/	NT	15	~2.5	720	3	63	75	20	2
Logan '91	Japan.	E /r/-/l/	NT	15	~10	4080	3	78	85	10	2
Pruitt '94	Engl.	H den-retr.	Lis & NT	30	~2	840	1	56	84	49	-17
Yamada '93	Japan.	E /r/-/l/	NT	45	~25	1224	9	70	89	28	-2
Flege '95	Mand.	E /t/-/d/	CD	7	~3.5	1680	3	66	77	16	-8
Flege '95	Mand.	E /t/-/d/	NT	7	~3.5	1690	3	67	83	24	-7

1. The Fading Technique. The first study demonstrating generalization beyond the training situation used synthetic stimuli to train native speakers of Canadian French to hear the English /θ/-/ð/ distinction [1]. Subjects were asked to identify each of a sequence of sounds as containing voiced or voiceless "th". Subjects received accuracy feedback immediately after each response.

Speech synthesis was used to create a sequence of signals, varying systematically in the amount of voiced or voiceless frication. At the start of training, listeners heard just two signals, one containing an exaggerated amount of the voiced target frication, and the other containing an exaggerated amount of the voiceless target frication. These "super-fricative" signals were designed to help the subjects attend to the target contrast immediately, and without making errors. As training progressed, additional signals with reduced amounts of frication were included in the training set, so that subjects gained experience with signals having more typical amounts of frication.

This approach improved the identification of tokens of natural speech; training with synthetic speech modelled on a male speaker generalized to natural tokens spoken by women as well as by men. Training with only a pair of "prototypical" speech signals was less effective than training with the full set of synthetic signals [16]. Moreover, training in word-initial position did not transfer to tokens containing the target sounds in other positions, nor to a task requiring subjects to identify sounds as containing one of *four* possible target sounds -- /d/ and /t/ and well as /θ/-/ð/ [17].

Targets in syllable-medial position contain cues common to those in word-final and word-initial position. However, training with syllable-medial tokens did not transfer substantially to target sounds occurring in word-final or word-initial position [18]. Thus, while fading with synthetic signals can improve perceptual ability substantially, learning seems specific to the phonetic environment in which the training sounds occurred.

2. Multiple natural tokens. Another successful technique for training new speech contrasts in adults involves the identification of multiple natural tokens of the target sounds. The first successful use of this technique used tokens from several talkers to train native speakers of Japanese to hear the English /r/-/l/ distinction [2].

Listeners identified each of a sequence of these tokens as being one of two words from a minimal pair, one containing an "r" and the other an "l". Incorrect responses were indicated by illuminating a light associated with the correct response, and then repeating the stimulus. Three weeks of such training improved identification performance by approximately 10%. When training used several talkers, learning generalized to novel words produced by a familiar talker and to a lesser extent to novel words produced by an unfamiliar talker (~8%; [2]). When training used a single talker, learning did not generalize to novel words produced by an unfamiliar talker [15]. Extending training to 9 weeks further improved performance [19].

The basic approach and stimulus set from [2] have also been applied to improve English /r/-/l/ identification for native speakers of the Korean language [20]. Training improved performance substantially for young Koreans, who had recently arrived in Canada; older Koreans who immigrated to Canada as adults received less benefit.

3. Alternating Listening & Identification Sets. A third successful approach was used to train English-speaking adults to identify Hindi dental and retroflex consonants [14]. This task cycled between sets of 50" listening trials" using a sequence of natural tokens of either dental or retroflex consonants, and sets of ten identification trials, each containing two repetitions of one of several possible dental or retroflex consonant tokens, followed by the subject's identification response, followed by accuracy feedback. This combination of listening and identification sets was repeated six times on each training day. The type of sound presented during each of the listening blocks was selected by the subject immediately prior to the block. Just one week of such training (~2 hrs) produced a 48% increase in average identification accuracy. Training also transferred to new words spoken by a new talker (with a 30% improvement from pretest performance).

This performance improvement is impressive, and it strongly encourages further work using this approach. Further research is required to evaluate the separate effects of, and interaction between, the listening task and the identification training task, used in [14].

4. Categorial Discrimination. Requiring listeners to categorize sounds in terms of the phonetic categories of the L2 is thought to be critically important for successful transfer of learning beyond the

training task. Normally, this focus on categorization is achieved by using an identification task. However, a **categorial** discrimination task (CD [21]) in which listeners must decide whether or not a pair of acoustically different signals are members of the same L2 phonetic category also focuses attention on linguistically-relevant categories. Thus, the CD task may also be effective for training new contrasts [22].

Recently, CD training was evaluated with native speakers of Mandarin for whom perception of unaspirated, English-language, word-final /t/ and /d/ is difficult [23]. Some subjects were trained with an identification task using multiple tokens of natural speech. Other subjects were trained in a categorial same/different task, in which two *different* tokens were presented on each trial, with the listener being required to indicate whether both were tokens from the same English-language category or whether the two tokens were from different English-language categories.

Three weeks of identification training improved identification accuracy by about 24%, while an equivalent amount of categorial discrimination training improved identification accuracy about 16%. Both types of training generalized to new tokens produced by a novel speaker; however, categorial discrimination training showed better retention, so that the performance of subjects from the two groups was more equivalent after a 2 month period without further training.

Training to Improve Speech Production.

Data relating systematic training to improved speech production abilities are presently very limited with adult L2 learners. However, training studies with young children who have difficulty producing sounds in their native language are encouraging. Many such children have correlated perceptual difficulties that identification training [24,25] improves.

These studies trained children in a "Category Inclusion" task, with feedback. Each child heard a series of speech sounds related to a specific production error manifested by the child. For each sound presented, the child indicated whether or not the sounds belonged to the target L1 category. Thus, a child who misarticulated /s/, would hear a sequence of misarticulated and correctly-produced utterances containing a target word containing /s/. For example, such a child could hear the word "shoe", spoken correctly and incorrectly, by many different speakers. The child touched a cartoon picture of a shoe, if a token was judged to have been pronounced correctly or a cartoon "X" when a token was judged to have been pronounced incorrectly.

The Category Inclusion task requires the subject to make an explicit judgement about whether or not a sound is appropriate for a particular linguistic category. Many more "inappropriate" sounds are included than in the standard "forced-choice" identification procedure. Importantly, such perceptual training not only improves children's identification performance, but it also transfers to speech production performance. However, this technique has not yet been evaluated with adult L2 learners.

Summary and Conclusions

When adults are trained using appropriate protocols, their abilities to perceive non-native speech contrasts can improve substantially. Appropriate protocols have the following characteristics: 1) they induce the learner to attend to cues relevant to assigning speech sounds to a phonemic category in the L2; 2) they provide prompt and unambiguous information about the appropriate categorization of each speech signal; 3) they use a set of speech tokens containing sufficient variability to permit subjects to learn about acoustic cues that are relevant to defining category membership in the L2 -- both cues relevant to inclusion of signals and cues irrelevant to classification in L2.

At least four training techniques that meet these criteria have now been demonstrated to produce effective training. All four can lead to improvements in speech perception that generalize beyond the training situation.

Such improvements have the following characteristics: 1) learning occurs quickly; 2) there is at least some transfer of training to the identification of novel (untrained) words, novel talkers, and novel phonetic environments; 3) there is minimal or no transfer to production; and 4) individual L2 listeners differ greatly in how much they benefit from training, even when age, linguistic background, and other factors are considered.

These results confirm that even in mature humans, those portions of the auditory system that are required to perceive and identify speech sounds remain relatively plastic. The speed with which such sizeable performance changes can be acquired suggests that training redirects attention more than inducing fundamental auditory system changes.

Optimizing Training: Directions for Future Research

For the L2 learner, the achievement of fluent, unaccented conversational speech is the ultimate objective. Performance clearly falls far short of this objective, even with extended training. However, there is much more reason to be optimistic, than there was a decade ago, when many speech researchers had reached the gloomy conclusion that adults had rather limited opportunities to learn new L2 contrasts because of permanent and irreversible neural changes.

Such a conclusion is no longer viable. However, it seems likely that we are still well short of what may be able to be achieved through systematic training. As impressive as the demonstrations reviewed here may seem, much can be done to further optimize training procedures. Answers to several questions are still required:

1). **How are listeners' L2 difficulties related to their L1 backgrounds?** Already, there is a substantial and growing body of work directed towards improving our understanding of this question. Good empirical work on this topic is very time consuming, but such work is continuing to appear [3,10]. Two significant theoretical positions which recently appeared on this topic [4,7] have helped to consolidate our understanding of empirical results. As this understanding improves, protocols for helping specific L1 groups to acquire specific L2 contrasts can be expected to become more effective.

2). **How can the differences between individual listeners, who have apparently similar linguistic backgrounds, auditory capabilities, etc., be understood?** This topic remains a challenge for many areas of speech perception. A particular question is how training can be personalized, through improved assessments of how a particular listener uses cues in the L1, and better targeting of the training stimuli and task for each listener. A relevant approach is provided in [26].

3) **How can protocols be structured to optimize learning and retention?** L2 training research remains largely in the "demonstration" phase, having an emphasis on determining whether or not some technique helps some L1 group to acquire some L2 contrast. Few studies have compared alternative training approaches and little protocol refinement has yet been attempted. Work may now be approaching a point of consolidation and refinement [8].

One very positive step has been the use of the stimulus set and training approach from [2] in several laboratories and across different subject populations.

Such sharing facilitates comparison across studies and should lead to more rapid advancement of knowledge.

4) **How should training be structured to optimize production ability?** The transfer of perceptual training to speech production has received such limited attention that few conclusions can be drawn. However, results with young children [24,25] suggest that such transfer may well be possible.

It seems unlikely that an "optimal" protocol can be created from a single one of the approaches examined to date, or indeed that any one protocol will be optimal for all listeners and situations. Rather, protocols will need to target subject needs, and components of each of several protocols may be used briefly at different points in time. For example, at the beginning of training, perceptual fading with synthetic signals may be used to direct the listener's attention to specific acoustic cues, without allowing others to vary [1]. Training with a structured sequence of natural tokens in a single phonetic context may then help the listener to classify sounds appropriately while ignoring irrelevant, naturally-occurring variation. The categorical inclusion task may then help the listener to further refine the L2 category. Training in additional phonetic environments may be appropriate. Finally, categorial discrimination may be used to consolidate learning and improve retention over the longer term.

ACKNOWLEDGEMENTS

*Supported by grants from NSERC and Starkey Canada. Direct e-mail to jamieson@audio.hhcru.uwo.ca.

REFERENCES 205-215.

[1]Jamieson, D.G. & Morosan, D.E. (1986), "Training non-native speech contrasts in adults: Acquisition of the English /θ/-/ð/ contrast by Francophones", *Perception & Psychophysics, 40(4)*, 205-215.

[2]Logan, J.S., Lively, S.E., & Pisoni, D.B. (1991), "Training Japanese listeners to identify English /r/ and /l/: A first report", *Journal of the Acoustical Society of America, 89*, 884-86.

[3]Flege, J.E., Munro, M.J. & MacKay, I.R.A. (1995), "Effects of age of second-language learning on the production of English consonants", *Speech Communication, 16*, 1-26.

[4]Best, C.T., McRoberts, G.W., & Sithole, N.M. (1988), "Examination of perceptual reorganization for non-native speech contrasts: Zulu click discriminations by English-speaking adults and infants", *Journal of Experimental Psychology: Human Perception and Performance, 14(3)*, 345-360.

[5]Flege, J.E. (1992), Speech learning in a second language. In Ferguson, C.A., Menn, L. & Stoel-Gammon, C. (Eds.) *Phonological development: Models, research, implications*, Timonium, MD: York Press.

[6]MacKain, K.S., Best, C.T., & Strange, W. (1981), "Categorical perception of English /r/ and /l/ by Japanese bilinguals", *Applied Psycholinguistics, 2(4)*, 369-390.

[7]Flege, J.E. (1991), Speech learning in a second language. In Ferguson, D., Mann, L., and Stoel-Gammon, C. (Eds.) *Phonological Development: Models, Research, and Application*, Parkton, MD: York Press.

[8]Rochet, B.L. (1994), "The efficient use of the computer in L2 pronunciation instruction", In Proceedings of the CALICO 1994 Annual Symposium on Human Factors, 178-182.

[9] Takata, Y. & Nábělek, A.K. (1990), "English consonant recognition in noise and in reverberation by Japanese and American listeners", *Journal of the Acoustical Society of America, 88(2)*, 663-666.

[10]Yamada, R.A. (1995), Age and acquisition of second language speech sounds: Perception of American English /r/ and /l/ by native speakers of Japanese.

In W. Strange (Ed.) *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research*, Timonium, (in press).

[11]Yamada, R.A. & Tohkura, Y. (1992), "The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners", *Perception & Psychophysics, 52(4)*, 376-392.

[12] Best, C.T. (1992), The emergence of language-specific phonemic influences in infant speech perception. In Nusbaum, H.C. & Goodman, J. (Eds.) *The transition from speech sounds to spoken words: the development of speech perception.* Cambridge, MA: MIT Press.

[13]Werker, J.F. and Pegg, J.E. (1992), Infant speech perception and phonological acquisition. In C.A. Ferguson, L. Menn, and C. Stoel-Gammon (Eds.) *Phonological Development: Models, Research Implications*, Timonium, MD: York Press. (pp 285-311).

[14]Pruitt, J.S. (1994), "Identification of Hindi dental and retroflex consonants by native English and Japanese speakers", *Journal of the Acoustical Society of America, 95(5) pt 2*, 3011.

[15]Lively, S.E., Logan, J.S., & Pisoni, D.B. (1993), "Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories", *Journal of the Acoustical Society of America, 94*, 1242-1255.

[16] Jamieson, D.G. & Morosan, D.E. (1989), "Training new, non-native speech contrasts: A comparison of two techniques", *Canadian Journal of Psychology, 43(1)*, 88-96.

[17]Morosan, D.E. & Jamieson, D.G. (1989), "Evaluation of a technique for training new speech contrasts: Generaliztion across voices, but not word-position or task", *Journal of Speech and Hearing Research, 32*, 501-511.

[18]Jamieson, D.G. & Moore, A.E. (1991), "Generalization of new speech contrasts trained using the fading technique", *XII Internation congress of Phonetic Sciences, 5*, 286-289.

[19]Yamada, R.A. (1993), "Effect of extended training on /r/ and /l/ identification by native speakers of Japanese", *Journal of the Acoustical Society of America, 93*, 2391.

[20]Jamieson, D.G. and Yu, K. "Perception of English /r/ and /l/ by adult native speakers of Korean", (unpublished)

[21]Polka, L. (1991), "Cross-language speech perception in adults: Phonemic, phonetic, and acoustic contributions", *Journal of the Acoustical Society of America, 89(6)*, 2961-2977.

[22]Strange, W. (1994), "Speech perception by second language learners" *Journal of the Acoustical Society of America, 95(5)* pt 2, 2998.

[23]Flege, J.E. (1995), "Two procedures for training a novel second-language phonetic contrast", *Applied Psycholinguistics* (in press).

[24] Rvachew, S. & Jamieson, D.G. Learning new speech contrasts: Evidence from adults learning a second language and children with speech disorder. In Strange ,W. (Ed.) *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research.* Timonium, MD: York Press (in press).

[25] Jamieson, D.G. & Rvachew, S. (1994), "Perception, production and training of new consonant contrasts with children having a functional articulation disorder", *Proceedings of the Third International Conference on Spoken Language Processing*, Yokohama: Acoustical Society of Japan, 1199-1202.

[26] Best, C.T. and Strange, W. (1992), "Effects of phonological and phonetic factors on cross-language perception of approximants", *Journal of Phonetics, 20*, 305-330.

PHONOLOGICAL AND PHONETIC ANALYSIS OF CLEFT PALATE SPEECH

Bernhardt, B. & Doan, A., University of British Columbia, Canada
Stoel-Gammon, C., University of Washington, USA

ABSTRACT

Speakers with repaired cleft palates may continue to have atypical speech development. Speech pathologists are often asked to comment on the need for, and effectiveness of, followup medical or prosthetic interventions. Few clinicians have access to technology for such determinations. We present articulatory-phonetic and nonlinear phonological analytic procedures which can assist in these determinations, using data from a child on a speech bulb reduction program.

INTRODUCTION

Although surgical repairs have become very effective for cleft palates in recent years, over 30% of persons with repaired cleft palate persist in having atypical speech development. Developmental substitutions (e.g. use of glides for liquids), compensatory substitutions (e.g. palatals, glottals, pharyngeals, or nasal snorts), and/or imbalances between oral and nasal resonance may continue to reduce speech intelligibility. A number of perpetuating factors may result in this persistent speech disorder. Surgeries may have been only moderately successful in achieving palate closure and lengthening. Chronic otitis media may have affected a child's perceptual bases for speech production. General developmental delay, ill health, or social interaction difficulties may have had a general negative impact on communication development. Even if none of these factors exist, habituation to the original characteristics of the oral mechanism (both structurally and functionally) can result in adherence to the original (deviant) phonological and phonetic system.

As part of the cleft palate team, speech-language pathologists, are not only responsible for direct speech intervention, but also are asked to make judgments as to the need for further surgical or prosthetic intervention, or the effectiveness of such procedures. Many useful technologies are currently available to assist the clinician in making such judgments. The velar port functioning can be assessed with tools such as multiview videofluoroscopy, nasoendoscopy, or nasometry. The speech signal can be evaluated through acoustic analysis. Electropalatography can inform about the placement of the tongue with respect to the palate. In selected hospital centers, such technologies have become useful adjuncts to perceptual judgments and analyses. However, many clinicians not working in (well-funded) hospital settings do not have such technologies available. They need to rely on phonetic transcription, general intelligibility assessment, and phonetic and phonological analyses to determine needs and effectiveness. As Howard (1993) [1] shows, detailed phonetic transcription and phonological analysis can lead to the judgment that a person with a cleft palate may have a well-developed phonological system with particular deficits in the articulatory realization of the phonological contrasts.

Inherent in this dichotomous interpretation of speech production is the assumption that intent and phonological representation are separate from the actual phonetic implementation of a word. In terms of phonology, we cannot confirm intent, nor determine the nature of underlying representation. All we can do is make inferences about a person's phonological system based on observation of *contrasts* and/or systematic sound *patterns* or *changes* in *phonetic* output. The more coherent and psychologically real our phonological theory, the more reliable and valid the inferences we make. In this paper we utilize aspects of current nonlinear (multilinear) phonological analysis. This theoretical framework allows detailed determination of patterns at various levels of representation: from the word, foot and syllable levels, to subsegmental featural levels (see below for elaboration). We also assume that phonological constraints operate on output, and that phonological constraints are often phonetically grounded (or motivated), following Archangeli & Pulleyblank (1994) [2]. Thus, the interrelationships between phonetics and phonology are assumed to be very close.

Phonological interpretation depends crucially on reliable and detailed phonetic transcription. The end-product of speech production processing is the articulatory-phonetic output. The more narrowly a listener is able to transcribe that output reliably, the more valid the data. Transcription skill aside, phonetic data tends to be generally "noisy," with even random variability with respect to exact location and timing of discrete phonetic units. A speaker with a repaired cleft palate may have a tendency to variability, as she or he accommodates to both developmental and induced mechanism changes, in the context of possible fluctuating hearing acuity or other perpetuating/predisposing factors. There may or may not be an attempt to reduce variability through more or less rigid adherence to the patterns of the early-established phonological and phonetic systems. Quantification of degree and type of variability is thus a relevant aspect of analysis.

The final determination of relative intactness of phonological and phonetic systems requires a balance between articulatory-phonetic and phonological analyses. This can be approached by attributing sufficient emphasis to phonetic detail and variability, while abstracting away from detail to infer general patterns. In this paper we will outline aspects of nonlinear phonological frameworks that lead to a detailed analysis of phonological systems, and utilize phonetic data in a variety of speech sample conditions to observe variability and consistency. The child whose analysis we use as an example had a speech bulb, the effectiveness of which was being evaluated one month after she started to wear it. We will utilize the nonlinear and phonetic analyses to outline similarities and differences with the speech bulb in and out, and between single words and connected speech.

SUBJECT

Tia (a pseudonym) was 5;11 at the time of this evaluation. She was born with a bilateral cleft lip and palate. Lip repair was done at 4 months of age, and initial palatal repair at 22 months of age. At the time of the palatoplasty, she also had a bilateral myringotomy to reduce otitis media. At 4;5 she had an orticochea pharyngoplasty to assist closure of the velopharyngeal port. (Orticochea pharyngoplasty is a type of pharyngoplasty in which a sphincter is created using the lateral and posterior pharyngeal walls.)

She and her family have received speech and language counselling or intervention services since she was an infant. Until about 3;6, she had recurrent otitis media with a fluctuating mild conductive hearing loss, a moderate delay in language comprehension, and a severe delay in language production. At the time of assessment, hearing status and language skills were within normal limits.

However, speech was and continues to be moderately unintelligible.

Her speech is characterized by hypernasality, nasal air emission and turbulence, and a moderate degree of use of compensatory and developmental phonological/phonetic patterns. Most frequent compensatory substitutions are glottals and pharyngeals, with some use of palatals for lingual consonants, nasals for oral stops, and nasal snorts. The glottal and pharyngeal substitutions are sometimes doubly articulated with oral place consonants. The most prevalent developmental pattern is the use of glides for liquids. Other developmental patterns include some word-initial and word-final omissions (for both singletons and clusters) assimilations, and occasional use of stops for fricatives. Generally, she matches the adult target more often word initially than in other word positions.

A nasoendoscopy after the orticochea pharyngoplasty revealed little movement of the palate or lateral pharyngeal walls. (Nasometry revealed an oral-nasal resonance imbalance, although we will not focus on the resonance issues in this paper.) Because of the velopharyngeal incompetence, and limited success of speech therapy in reducing the deviant speech patterns, the cleft palate team decided to implement a speech bulb reduction program.

PROCEDURES

Speech bulb program

A speech bulb is a dental appliance with a velar extension which closes off the velar port maximally (although as we see for Tia, some nasal emission still occurred, meaning complete occlusion was not achieved). Speech therapy continues crucially after the child receives the bulb in order to eliminate compensatory patterns while the mechanism is being normalized through use of the appliance. At appropriate intervals, the speech bulb is reduced minimally in size, with the hope that the child will initiate use of the lateral pharyngeal walls to achieve closure, in order to continue to have a normalized mechanism. Before such reductions are made, however, notable improvement in speech production needs to occur.

The data

Tia had worn her speech bulb for about a month when the videotaped data for this analysis were collected (a typical period for a post-bulb evaluation). In her case, three sets of data were available for the evaluation analysis: a small set of data with the bulb out, and a longer data set with the bulb in, both in single words and connected speech.Sufficient pre-bulb data and speech bulb out data were not available from the hospital to make a complete pre-post comparison. However, sufficient data are available to comment on the variable effects of the bulb in single words and connected speech. This paper is not an evaluation of the speech bulb program for this child, but an example of analytic procedures. We would of course recommend the same type of pre-analysis for clinical evaluation of the speech bulb for a given client.

The general analysis framework

Following phonetic transcription of the three data sets (speech bulb out, and speech bulb in, single words versus connected speech), data was analyzed in four ways. Syllable and word structure characteristics were defined, and three types of feature description were made, two of which related to nonlinear feature geometry, and one to actual phonetic output. Proportional matches with the adult target were calculated for each of the nonlinear codings, and cumulatively across all three feature conditions for phonetic coding, the rationale being that phonetic realization is the end-product of phonological processing, and therefore

subsumes all previous nonmatches. We elaborate on the nonlinear descriptions below.

Nonlinear analysis

Nonlinear phonological theory focuses on the hierarchical nature of relationships among phonological elements. Two major levels of phonological organization are associated in principled ways: the prosodic level and the segmental level. The prosodic level includes all structure above the level of the segment and ultimately ties in with stress and intonation patterns. The lowest level of prosodic structure is subsyllabic structure, i.e., the onset or rhyme ([b] onset versus [æt] rhyme of bat). Progressively larger units of structure include the syllable, the foot (incorporating strong and weak syllables), the word, and ultimately, the phrase. The segmental level is described in most accounts as an hierarchy of features, each of which has some autonomy within the constraints of the hierarchical relationships. At the highest level of the hierarchy, the Root Node is the sum total of all of the features and this node connects to the prosodic tiers "above." The features immediately dominated by the Root Node are the manner and sound class features. In this account, following Bernhardt & Stoel-Gammon (1994) [3], we use:

a) [+consonantal] to refer to true consonants
b) [+sonorant] to refer to vowels and glides
c) [+continuant] to refer distinctively to fricatives (including /h/)
d) [+consonantal]-[+sonorant] to refer to liquids
e) [-continuant] to refer to stops, and
f) [+nasal] to refer distinctively to nasals.

These categories were sufficient for the speaker in question.

Also dominated by the Root Node are the Laryngeal Nodes and Place Nodes, which dominate their own set of features. Laryngeal Node features used in this study are:

a) [+voice] and [-voice]
b) [+spread glottis] for /h/
c) [+constricted glottis] for [ʔ]

Place Node features used in this paper required some elaboration beyond what is typically needed for English, due to the compensatory substitutions involving place:

a) Labial, referring to all labial articulations, including /p/, /b/, /m/, /w/, /f/, /v/, and /ɹ/ (the latter in combination with Coronal)
b) Coronal, referring to all consonants using the tip and blade of the tongue, and including /t/, /d/, /s/, /z/, /ʃ/, /ʒ/, /tʃ/, /dʒ/, /θ/, /ð/, plus /j/ and other palatals in combination with Dorsal
c) Coronal [-anterior], referring only to blade articulations: /ʃ/, /ʒ/, /tʃ/, /dʒ/
d) Coronal [+distributed], referring to interdentals /θ/ and /ð/
e) Dorsal, referring to articulations with the body of the tongue and pharynx. This includes both the typical dorsals of English (/k/, /g/, and /ŋ/) which are [+high] and uvulars and pharyngeals which are [+low], two of her compensatory substitutions.
f) Place combinations: Labial-Coronal, for /ɹ/, Dorsal-Coronal for palatals (also involved in compensatory substitutions), and [+constricted glottis]-Other Place for doubly articulated consonants with glottal and other places, Dorsal-[+low]-Other Place for consonants with both pharyngeal or uvular and other places of articulation.

Another subtheory of nonlinear phonological theory important to the feature designation used here is the theory of underspecification. According to this minimalist theory, only the unpredictable

features of a segment are present in underlying representation, other predictable features assumed to be encoded during processing. In phonetic output, then, more features are assumed to be present than underlylingly, but even in output, a minimalist description of relevant distinctive features is assumed. Thus,when we describe the phonetic output for labial stops, we will only refer to the minimum number of features which distinguish them unless we have reason to include others: [-continuant], [voice] features, and [Labial]. Where categories overlap, because of phonetic implementation or phonological aberrations, additional features will be added. Thus, if the child is having trouble differentiating nasals and stops, the feature [-nasal] may become a relevant comment for that child on labial stop production, even though it would not normally be needed in phonological and phonetic description.

As noted above, levels of representation have some autonomy in terms of phonological constraints and processes, but also interact with one another in principled ways. Features are combined into segments, and segments are combinations of features, and can be viewed componentially. To exemplify:

a) The feature [Labial] may be established in a phonological system, but not in combination with all manner or voice features. Thus, /m/ and /w/ may be possible segments, but /f/ and /b/ may not be possible segments, surfacing as /m/ or voiceless [m]. Thus, the feature [Labial] is intact, but it can only cooccur with [+nasal], and not with [+continuant] or [-nasal].

b) Alternately, the feature [Labial] may not occur at all: /m/, /b/, etc., all surfacing with some other place of articulation.

c) The feature [+continuant] may be present in the system and realized in combination with [+spread glottis] for /h/, but may not cooccur with oral Place features, which all surface as [h]. Thus, the feature [+continuant] is present in output, but not in combination with other features.

d) Alternately, no fricatives, either oral or glottal may appear, the feature [+continuant] thus not occurring at all.

e) There may be voicing constraints. For example, a child may produce /k/ for both /k/ and /g/, but produce /t/ and /d/. When this is the case, the features [Dorsal] and [-continuant] are present, but the feature [+voice] does not cooccur in output with the [Dorsal] and [-continuant] features, only in combination with [Coronal] and [-continuant].

f) Alternately, no [+voice] obstruents may occur, in which case the feature [+voice] is not yet established in the system.

These constraints on feature presence or cooccurrence may apply across word positions, or only to one word position. In any event, we are able to describe the output data phonologically in such ways, deriving a perspective on major and minor problem areas for a child: whether in terms of feature establishment per se, or in terms of feature cooccurrence.

Final phonetic product

In the analysis we present, phonological features are examined in terms of their presence and their combinatorial power in the various speech conditions. However, the end-product of phonological processing is the articulatory-phonetic realization. Thus, both the phonological mismatch types are subsumed in this last category, and further details of phonetic deviance are included. Note that we are not in any of these cases making inferences about phonological representation per se. We assume that phonological constraints

operate on output, and thus can be observable in the phonetic output. In the final production, the feature [Labial] may cooccur with the appropriate features, but have some component of difference which infers a phonetic realization problem. In other words, phonological contrast is evident, and features combined, but phonetic realization results in some overlap or imprecision. Note that in the features we described above, details of implementation are not generally included. For example, [Labial] does not distinguish between bilabials and labiodentals. In English, labial fricatives are labiodental, and hence that detail of phonetic implementation is not considered a phonologically relevant distinction, even though it is phonetically important. The [+anterior] coronals in English are alveolar in phonetic place of articulation, but in other languages [+anterior] coronals are dental, another example of a relevant phonetic and irrelevant phonological distinction. These differences are important in the examination of deviant speech, where mechanism abnormalities can result in an "irrelevant phonetic distinction" becoming both "phonologically" and "phonetically relevant." Examples of such end-product differences include:

a) Nasal emission superimposed on oral articulation, indicating phonological intent to produce an oral consonant, with insufficient velar port closure

b) Weak stop productions that have a fricative-like aspect, but that are still distinguishable from true fricatives in the child's output

b) Place of articulation that is within the region of the intended articulation (and phonologically "accurate") but is phonetically different. Examples of such deviations may include:

i) Labiodentals produced with the lower teeth and upper lip

ii) All labials produced with the teeth and the lips, but with consistent manner differences to distinguish them from each other and from the /f/ and /v/

iii) Coronals produced in the palatal region, provided they are consistently distinguished from dorsal target consonants

iv) Dorsals produced consistently as palatals or uvulars in a way that consistently distinguishes them from coronals

c) Voicing that distinguishes between voiced and unvoiced segments, but not with the expected adult target values of the language. For English, this may imply a failure to aspirate word-initial stops.

In summary, then, the data were examined in the three conditions: speech bulb out, and speech bulb in, single words and connected speech. The phonetic realizations were coded in terms of:

a) feature presence (using the features described above, and including the concept of underspecification)

b) feature cooccurrence, and

c) final phonetic output.

Tia's speech production: Syllable and word structure

As we commented earlier, syllable and word structure was reasonably well-developed. Thus, phonological constraints for word production were minimal, except with respect to consonant clusters, which were still developing. Across the conditions, she was able to produce up to 3-syllable words consistently. Word shape matches were better in the connected speech condition than the single-word condition. This will be seen to be a trend opposite to that for feature production, and may reflect the particular words sampled rather than a

better ability to produce clusters in running speech. In any event, phonological output at the prosodic structure level was adequate with the speech bulb in and out.

Tia's speech production: Features

There is not room in this paper for a complete examination of the features in terms of occurrence, cooccurrence and final phonetic realization. Hence, we will summarize the relevant procedures and results here, and present a table which exemplifies the analysis.

Feature occurrence

The major feature occurrences of concern were [continuant] (for stops and fricatives), [+consonantal]-[+sonorant] (for liquids), Coronal, and Dorsal. Tongue placement and oral pressure are implicated for all those but the liquids (which could have been developmental). Hence the speech bulb program should have had an impact on feature production. This was in fact the case for the Root (manner) features involving obstruents. In the 'bulb in-single words' condition, [-continuant] was 96% accurate in terms of occurrence, and [+continuant] was 85% accurate. Performance was somewhat worse in the connected speech condition (93% and 68% respectively), but this was an improvement over the 45% accuracy for [-continuant] in the bulb-out condition. Coronal and Dorsal were approximately the same across conditions for feature occurrence, which is important in the evaluation of the speech bulb. Compensatory articulations often affect place of articulation, and if they did not change noticeably in the month, she was not yet ready for bulb reduction, and it can be assumed that phonological habituation was strong for place. The artificial occlusion of the velar port, however, did increase her ability to produce high-pressure obstruents, indicating that phonologically, those features were well-established, and that mechanism was affecting phonetic realization.

Feature cooccurrence

Additional concerns can become apparent when examining feature cooccurrence. Either new features can show up as problematic, or features already of concern in terms of occurrence can decrease in accuracy. Cooccurrence accuracy implies that all Root, Laryngeal, and Place features of a segment are accurate, unless separate tabulation is made for Root-Laryngeal, Root-Place, and Laryngeal-Place combinations (which we have not done here in terms of brevity). Feature cooccurrence here is thus a measure of *segmental* phonological accuracy.

Decreases in accuracy were noted for [-continuant] and [Dorsal] across the 'bulb out' and 'bulb in-connected' speech conditions. For example, in the 'bulb out' condition, accuracy for [-continuant] decreased from 45% to 33%, and in the connected speech condition, from 93% to 77%. However, in the 'bulb in-single words' condition, accuracy was maintained for these features at a fairly high level. For example, [-continuant] was 96% and 95% across the two conditions. The differences between the single words and connected speech conditions for these features does indicate some continuing concern for speech bulb reduction however, particularly since the dorsal stop substitutions involved compensatory palatal and uvular articulations, and other manner errors involved nasal substitutions for stops.

No changes were noted for Coronal or liquids.

New features involved in the feature cooccurrence analysis were [Labial] and [+voice]. Hence, although the [Labial] and [voice] features were reasonably well-established in terms of occurrence, cooccurrence restrictions affected them also across conditions, although least in the 'bulb in-single words' condition.

Phonetic end-product

The final analysis was of the phonetic variants. As the sum total of *all* deviations, decreases in performance can be expected, but may not necessarily occur. Further decreases in accuracy were noted for:

a) [-continuant] in all conditions
b) [+continuant] in the 'bulb in' conditions
c) Labial and Coronal in 'bulb out' and Coronal in 'bulb-in' conditions
d) [-voice] in the 'bulb out' and connected speech conditions.

These tendencies further emphasize the phonetic difficulty with high-pressure obstruents and nasal emission, suggesting that occlusion was not sufficient with the particular bulb being used. Again, the place features continued to show deviations of nasalization and precision of placement, showing phonetic difficulty and compensation not yet corrected by the appliance and the therapy program. See Table 1.

CONCLUSION

An example of an analysis methodology differentiating types and degrees of phonological and phonetic features was presented for a child with a repaired cleft palate who was on a speech bulb reduction program. By dividing phonological features into occurrence and cooccurrence categories, and also separating out a phonetic category, relative robustness of features was identified. Coronal and Dorsal place were problematic across typologies and bulb-in/out conditions. Labial and [voice] features became implicated in the cooccurrence and phonetic conditions, showing they were less robust than phonological occurrence indicated. Furthermore, obstruent manner features responded positively to the speech bulb, although less so in connected speech. Overall the speech bulb program was having some influence on obstruent manner production (because of occlusion of the nasal cavity), but not sufficient influence in connected speech. Place and voice remained problematic, and hence the compensatory substitutions were not yet sufficiently diminished for reduction to be done. The methodology is thus seen to have potential for use in clinical situations, particularly when technological assessment is not available.

Table 1. Place Node Features: Bulb-in, connected speech

	Present?	Cooccurring?	Phonetic accuracy
Labial	95%	87%	84%
Coronal	79%	71%	63%
Dorsal	98%	77%	74%

REFERENCES

[1] Howard, S. 1993. Articulatory constraints on a phonological system. *Clinical Linguistic and Phonetics, 7*, 299-317.

[2] Archangeli, D. & Pulleyblank, D. 1994. *Grounded Phonology*. Cambridge, MA: MIT Press.

[3] Bernhardt, B. & Stoel-Gammon, C. 1994. Nonlinear phonology. *JSHR, 37*, 123-143.

THE CHARACTERIZATION OF DISORDERED CHILD PHONOLOGY

P. GRUNWELL

DeMontfort University, Leicester, U.K.

ABSTRACT

Phonetic and phonological studies of disordered child phonology are regularly reported in the clinical linguistic and speech pathological literature. The characterization of the disordered phonology is adult-centred and heavily influenced by current phonological theory. In consequence a clinical phonological assessment is an error analysis and phonological change is an all-or-none phenomenon. The basic phonetic and phonological concepts have yet to make an impact on the clinical investigation of child phonology.

INTRODUCTION

This presentation, like the two other contributions to this symposium, focuses on child language disorders, specifically developmental phonological disorder (or disability). A linguistic definition of this disorder is : "a linguistic disorder manifested by the use of abnormal patterns in the spoken medium of language" [1]. Many attempts have been made to characterize this disorder using analytical and assessment techniques derived from phonological theory. These descriptions have tended to be comparative error analyses that do not consider the children's pronunciation patterns as an independent phonological system. In consequence the clinical evaluation and management of the disorder do not take into account the dynamics of phonological functioning and phonological change.

DEVELOPMENTAL FRAMEWORK

Phonological process analysis has, since the early 1980s, become the dominant framework for the characterization of developmental phonological disorders. This approach has however become somewhat removed from the original theory upon which it is based, viz: Natural Phonology [2]. Stampe's theory is an attempt at an explanatory theory which is rooted in the phonetic bases of human language [3]. A substantial proportion of Stampe's evidence for his theory is child speech data. It is not doubt this fact which made his framework attractive to clinical phonologists. There is however a discontinuity between Stampe's theoretical exposition and the practical applications of his concepts in the clinical context. At the most basic level, many of the developmental phonological processes included in clinical assessment procedures such as Natural Process Analysis [4], Assessment of Phonological Processes [5] and Phonological Assessment of Child Speech [6] are not mentioned by Stampe by name nor even described by him.

Stampe's approach to phonology is also attractive clinically because his key theoretical concept, the phonological process is defined by him as a "phonological substitution". This concept therefore enables the traditional approach of error analysis to be presented in phonological terms. Process analysis focuses on the patterns in children's mispronunciations, with little regard to issues of naturalness. The processes are presented as developmental patterns and the characterization of disorder is therefore in a developmental framework.

The five characteristics of developmental phonological disorder in this framework are:

* persisting normal processes
* chronological mismatch
* unusual processes
* variable use of processes
* systematic sound preference

[6].

The first three of these are clearly developmental in nature being defined by what is expected to occur in normally developing phonologies. The last two are also developmental in definition but in addition refer to characteristics of the occurrence of processes in disordered child phonology. As such they could be viewed as phonological rather than developmental characteristics (see further below). Excluding these two, the developmental characteristics can be summed up as involving delayed, uneven and deviant phonological development, each type occurring in some measure in each case.

Unusual Processes

Within this framework attempts have been made to identify the characteristics of unusual processes, most notably by Leonard [7]. He identifies three types:

* Salient but unusual sound changes with readily detectable systematicity.
* Salient but unusual sound changes with less readily detectable systematicity.
* Subtle phonological behaviours.

The first type includes unusual substitutions defined developmentally, such as late sounds occurring before early sounds, phonologically, such as use of sounds absent from the model language and phonetically, such as use of sounds absent from natural language. The second type includes context-based substitution patterns such as assimilation, dissimilation, metathesis and consonant-vowel interactions. The third type includes children's pronunciations that are perceived by adult listeners as homophones eg tea and key as [ti] but which instrumental analysis (either spectrographic or electropalatographic) reveals to be different. This characterization of unusual processes goes beyond the deterministic view of phonological development presented by Natural Phonology and beyond the constraints of phonological process analysis as employed in clinical assessment procedures. It shifts the focus from the developmental dimension to the phonological .

PHONOLOGICAL FRAMEWORK

Within a phonological framework three parameters have been identified on which disordered phonological patterns can be described. These parameters are:

* system
* structure
* stability

These are the fundamental parameters of phonological organization and functioning. Phonologies operate with systems of contrastive units, within structural constraints determining the sequence and order of these units and in a state of relative stability or homeostasis.

In seeking to characterize disordered phonology it is necessary to examine the phonological patterns within these three parameters. In the evaluation of system , the size of the child's phonetic inventory and the child's contrastive system is analyzed by comparison with the adult targets attempted. In assessing structure, the distribution and combination of consonants and the syllabic structures in the child's pronunciation patterns are analyzed by comparison with the adult targets attempted. In regard to stability, the consistency or variability in the child's realizations of the adult targets is

examined.

Clinical research indicates that children's disordered phonologies can be definitively characterized using these three parameters. Disordered phonological systems are characteristically restricted and symmetrical. The asymmetry is evident in that potential contrastive feature combinations are not exploited. For example, this is the consonant inventory of one phonologically disordered English-speaking child:

m n
p b t g
f

This child's inventory contains all the possible feature-contrasts to create the adult system, viz:nasal/plosive/fricative; bibabial/alveolar/velar;voiced/voiceless; but it does not combine them to produce the maximum number of contrasts.

Disordered phonologies are structurally restricted in similar ways. Some children have markedly different consonant systems at different places in structure. For example the consonant distribution patterns for one English speaking phonologically disordered child were:
SIWI: m n b d g h w j
SFWF : m n p t f v s z.
Another example of a child having the potential to realize many adult targets but exploiting that potential. Further typical instances of structural asymmetry that have been frequently observed in phonologically disordered English-speaking children are the tendencies for one position in syllable and word structure to be relatively well-developed vis-a-vis the others. The most common tendency is for the range of contrasts in word initial position (SIWI) to be larger and more closely matched to the target system than in other word positions. Typically in within word positions (SIWW) there will be a very restricted range of consonants; often glottal consonants [ʔh] are dominant; and in word final position, (SFWF) zero realizations predominate, ie the open syllable is the canonical structure. Where English is the target language this type of patterning is seriously dysfunctional and communicatively inadequate; with other target languages, for example Italian, this would not be so.

The parameter of stability reflects the finding that disordered child phonologies tend to evidence variable realizations of the same adult targets. It is important to examine the occurrences of stability and variability in order to identify whether there is any latent potential to expand the system of contrasts. Logically, there are four types of consistency/variability:

* consistently correct match:
/t/ -> [t]

This so-called 'correct' realization may be compromised, however, if the same phone (i.e. [t]) is used to realize another target, e.g. /k/->[t]. In such a situation the child's [t] is not phonologically contrastive.

* consistently incorrect match:
/k/ -> [t]

This is likely to entail a lack of contrast as in the above example. However if one target is uniquely realized by a phone not in the target system,e.g. /s/ -> [n]. This would not result in phonological inadequacy.

* inconsistently correct match:
/k/ -> [k t]

The contrast is potentially present; this is progressive variability.

* inconsistently incorrect match:
/k/ -> [t d g]

There is no apparent potential for the contrast to develop. This variability does not appear to be progressive.

Using these three parameters to characterize disordered phonologies enables a phonological analysis and assessment which identifies the dynamics of phonological functioning and the potential for phonological change. Treatment aims are therefore defined by the characteristics of the child's phonology. They extend and expand on the potential within the child's patterns.

CONCLUSION

This paper has focused on the phonological characteristics of disordered child language in an attempt to define this type of developmental disability in linguistic terms. In this conclusion a number of other issues that are pertinent will be mentioned as items for further exploration and discussion.

This paper, following the vast majority of the literature, solely discusses the nature of consonantal patterns. In a review of recent studies, Grunwell [8] discusses the characteristics of disordered vowel patterns. Whilst at present there are relatively few studies, it would appear that vowel disorders also show phonological patterning in that (i) vowel errors shared features in common with the targets and (ii) the errors entail losses of contrasts. More evidence is required to investigate this area further.

As has already been mentioned some children signal contrasts imperceptibly. On the other hand some children use very unusual patterns for certain targets, such as /kl/ and /pl/ -> [ɬ]. These phenomena prompt the question : what is the phonetic basis of these realizations? It highlights the fact that children are abstracting perceptual information from the speech they hear and on the basis of this information attempting to create their own phonetic and phonological systems. It is apparent that in some instances children seem to attend to aspects of the speech signal that are not salient for mature speakers. This is another area which would profit from further investigation.

Finally clinical phonetics and phonology needs to investigate further the interaction between phonological development and the presence of a physiological impairment to the speech production mechanism, such as a repaired cleft palate. As reported by Grunwell [9] in many instances phonological development follows an essentially normal path in spite of the physical abnormality. Further investigations of other types of abnormalities are required to enhance our understanding of the nature of phonological development and phonological disorder.

REFERENCES

[1] Grunwell, P. (1981), *The nature of phonological disability in children*, London : Academic Press.
[2] Stampe, D. (1979), *A dissertation on natural phonology*, New York : Garland Publishing Inc.
[3] Donegan, P.J. and Stampe, D. (1979), The study of natural phonology, In D.A. Dinnsen (Ed.), *Current approaches to phonological theory*, Bloomington : Indiana University Press.
[4] Shriberg, L.D. and Kwiatkowski, J. (1980), *Natural process analysis*, New York : John Wiley.
[5] Hodson, B.W. (1980), *Assessment of phonological processes*, Danville, IL : Interstate Inc.
[6] Grunwell, P. (1985), *Phonological assessment of child speech*, Windsor : NFER-Nelson.
[7] Leonard, L.B. (1985), "Unusual and subtle behaviour in the speech of phonologically disordered children", *Journal of Speech and Hearing Disorders*, vol. 50, pp. 4-13.
[8] Grunwell, P. (1995), "Changing phonological patterns", *Child Language Teaching and Therapy*, vol. 11, pp. 61-78.
[9] Grunwell, P. (Ed.) (1993), *Analyzing cleft palate speech*, London : Whurr.

QUANTIFYING TIME-VARYING SPECTRA OF ENGLISH FRICATIVES

Lorin Wilde
Research Lab of Electronics, Dept. Electrical Engineering & Computer Science
MIT, Cambridge, MA 02139, USA
email: wilde@speech.mit.edu

ABSTRACT

Noise characteristics of fricatives were quantified with respect to adjacent vowel spectra. The weak and strong fricatives were well-separated: the maximum amplitudes above 2 kHz in the fricative, normalized relative to vowel amplitude, were 15-20 dB more for the alveolars and palato-alveolars than for the labiodentals and dentals. In addition, spectral changes during the consonantal interval were calculated.

1. INTRODUCTION

The acoustic consequences during fricative production include continuous spectral variations over time. First, the articulation and aerodynamics in noise generation during a particular fricative are continuous. In addition, the acoustics of fricatives produced in connected speech are influenced by concurrent coarticulatory movements.

Recent studies have provided additional evidence that the kinematics of fricative articulation create an acoustic signal that is inherently non-static [1] [2] [3]. Difficulty in the analysis of fricatives also arises from the nature of random noise generation in fricative production. The nature of a noise source complicates the accurate measurement of spectral properties associated with the articulatory movement.

An automatic analysis system for quantifying fricative noise spectra was developed. The objective was to reduce the dimensionality of the data while measuring essential spectral properties. Spectral changes during the consonant were examined. In addition, the attribute of stridency, signaled by greater energy in the high frequencies in the consonant relative to the vowel, was examined. The following questions motivated the choice of acoustic measures: 1) How much greater energy? and 2) In which frequency regions?

2. METHODS

A database was collected in order to examine in detail the acoustic attributes of fricative consonants in the front, back and back-rounded vowel contexts. Three normal speakers of American English, one male and two female, recorded 'CVCV'CVC nonsense syllables. The consonant was one of the eight English fricatives: /f, v, θ, ð, s, z, š, ž/ and the stressed vowels were /i, ɛ, ɑ, ʌ, o, u/. The first and third vowels in an utterance were the same. Two repetitions of each fricative in pre-stressed position were analyzed in this study.

The speech was recorded in a sound-treated room using an omnidirectional microphone which was located approximately 25 cm in front of the speaker and 5 cm above the speaker's mouth. The recordings were low-pass filtered at 7.5 kz and digitized at 16 kHz. One additional male speaker, previously low-pass filtered at 4.8 kHz and digitized at 10 kHz by Klatt [4], was also studied. The combined database was used to develop an automatic analysis system for quantifying fricative noise spectra.

Fricative noise characteristics were considered with respect to adjacent vowel spectra, with measures made relative to the consonant-vowel (CV) boundary. Digitized waveforms labeled with acoustic landmarks, i.e. fricative-vowel boundaries, are the inputs to the analysis system. Averaged spectrograms were computed by advancing a 6.4 msec window in 1 msec steps and averaging overlapping windows. A 20 ms averaging interval was empirically chosen: long enough to reduce error due to random fluctuations and short enough to quantify time variations in individual tokens. The maximum power was calculated in five frequency bands: 1) 0-0.5 kHz, 2) 0.5-1 kHz, 3) 1-2 kHz, 4) 2-4 kHz and 5) 4-8 kHz. The amplitudes and frequencies of spectral peaks, occurring relative to landmark times, are the outputs. Further details are provided in Wilde [5].

3. RESULTS

The following results are reported here: 1) variation in the noise over the duration of the consonant and 2) quantification of the feature [strident]. All measures are made relative to the consonant-vowel (CV) boundary. The following results are reported for the voiceless fricatives, in order to restrict our discussion to utterances for which the CV landmark could be accurately identified to within 4 ms.

3.1 Time-varying Noise Spectra

A measure of spectral variation over time was calculated by subtracting the amplitude value at the right edge of the fricative ($CV - 20$ ms) from the amplitude value at the temporal center of the fricative. A negative difference means that the amplitude at the edge is greater than the amplitude at the midpoint. The results for the three highest frequency bands are shown for all four speakers in Figure 1. The main, not unexpected finding is that there is considerable variation in noise spectra over time. That is, the noise amplitude is not constant and, from the interquartile ranges of all subjects, appears to vary from about -13 to +8 dB over the interval from the fricative midpoint to just before the fricative-vowel boundary.

The individual results for each band suggest a trend for differences between the weak and strong fricatives. For Band 3 (1-2 kHz) the clear trend is that there is greater amplitude difference for the labiodental and dental fricatives, grouped here as nonstrident. Band 4 (2-4 kHz) shows the same trend, although the ranges are more similar. In the 1-4 kHz range, the differences for all fricatives are negative, i.e., the edge is stronger than the middle. However, for the highest frequencies in Band 5 (4-8 kHz) there is a contrast in trends between the nonstri-

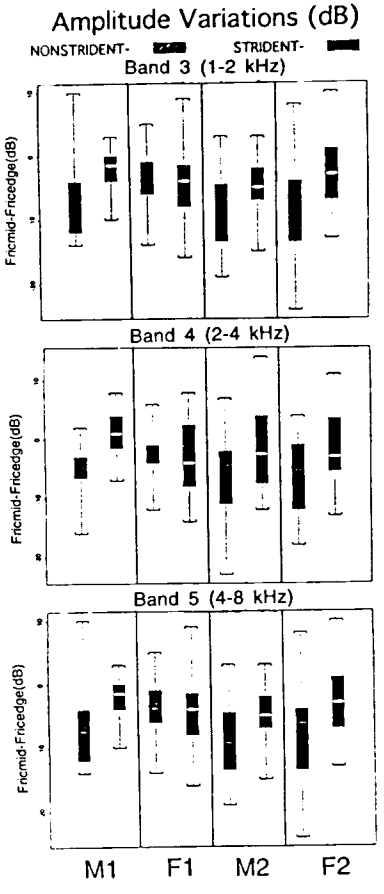


Figure 1: *The medians (lines) and interquartiles ranges (IQR=box height) illustrate the magnitude of amplitude variations (dB) in Bands 3 (top), 4 (middle) and 5 (bottom). Each box represents 24 data points, each calculated by subtracting the amplitude at the right edge (CV − 20) from the amplitude in the middle (CV − duration/2), with nonstrident vs. strident voiceless fricatives are shown separately for each speaker. The dotted lines extend to the extreme values of the data or a distance 1.5xIQR from the center, whichever is less.*

dents, which are clearly negative, and the stridents, which are clearly positive.

3.2 Quantifying Stridency

The feature [strident] was quantified by subtracting the F1 amplitude in the following vowel from the maximum amplitude peak above 2000 Hz in the consonant. As expected, the weak fricatives are well-separated from the strong fricatives. For example, the mean amplitude differences between /θ/ and /s/, which have the closest relative location of supraglottal constrictions, range from 13.7 to 19.9 dB for measures made at fricative midpoint and from 12 to 21.5 dB for measures made at the right edge of the fricative.

We can also compare these normalized amplitudes averaged separately for the weak voiceless fricatives (/f, θ/) and for the strong voiceless fricatives (/s, š/), which we have grouped as nonstrident and strident fricatives, respectively. The average normalized amplitudes for nonstrident and strident fricatives, measured at the edge of the fricative (at relative time = $CV - 20$) and normalized with respect to F1 amplitude (at relative time = $CV + 20$) are shown in Table 1. The difference between the grand average means for nonstrident and strident fricatives, computed as the average of the means of individual subjects, is 17 dB.

Table 1: *Means and standard deviations of normalized amplitudes (dB) for the nonstrident and strident fricatives for all four subjects. The normalized amplitudes were found by subtracting the amplitude of the first formant (at $CV + 20$ ms) from the maximum peak above 2000 Hz (at $CV - 20$ ms).*

Normalized Amplitudes (dB)			
Speaker		Means	S.D.
F1	Nonstrident	-41	4.74
	Strident	-23	5.58
F2	Nonstrident	-31	3.88
	Strident	-16	5.46
M1	Nonstrident	-33	3.71
	Strident	-18	4.93
M2	Nonstrident	-41	8.11
	Strident	-19	5.31
Average	Nonstrident	-36	7.07
	Strident	-19	5.90

4. DISCUSSION

In quantifying the time-varying spectra of fricatives, we asked the following questions: How big a change and in which frequency regions? The observed amplitude changes of -13 to +8 dB from the fricative midpoint to the fricative-vowel boundary exceed the expected error from the noise source, and presumably reflect movements of the major articulators in forming and releasing the supraglottal constriction.

The edges were stronger than the middle for all fricatives in the mid-frequency bands (1-4 kHz), consistent with observed excitation of the second and third formants near the vowel boundary, and with the presence of aspiration in the vicinity of the fricative-vowel boundary. It should be noted that back cavity excitation can also reflect incomplete pole-zero cancellation which can occur when there is coupling between the front and back cavities. Often, there is a short (less than 20 ms) gap, where neither the frication nor aspiration noise is very strong. A short gap in energy at the boundary between a voiceless fricative and the following vowel could be interpreted as reflecting that the supraglottal constriction is released before the glottis is closed. Presumably this reflects the mistiming between turning off the noise source for the fricative and turning on the voicing source for the following vowel.

Significant spectrum amplitude differences were observed at higher frequencies (4-8 kHz) between the nonstrident and strident fricatives. For the strident fricatives, the highest frequencies are strongest in the middle of the consonant, when the cross-sectional area of the supraglottal constriction may reach its minimum. This finding is consistent with a previous study [1], an LPC analysis of voiceless fricatives preceding five vowels, in which high-frequency peaks to tended to appear more often in the midpoint of a fricative than in the initial or final 15 ms.

The nonstrident fricatives in English show greater overall variability in amplitude than the stridents. Results of Utman and Blumstein [6] suggest that the realization of an acoustic property is influenced by the linguistic role its associated feature plays in a particular language's sound inventory.

The normalized amplitudes of the weak and strong fricatives in English were well-separated: the maximum amplitude above 2 kHz in the fricative, normalized relative to vowel amplitude, is 15-20 dB more for /s/ and /š/ than for /f/ and /θ/. Spectral differences between strident and nonstrident fricatives suggested that models of the filtering of the noise source by the front cavity might be improved if the losses in the vocal tract were better represented and if better estimates could be made of the source location.

5. SUMMARY

In the present analysis, the amplitudes in restricted frequency regions of fricative noise were examined with respect to the neighboring vowel. It was hypothesized that relative measures could be found to capture important characteristics of the time-varying noise and reduce the dimensionality of the data. Studying noisy speech sounds yields inherently noisy findings. We observe noise variations over time, variations from one token to another and inter-speaker variability. Our calculations of the amplitude variations in selected frequency bands for English fricatives guide understanding of the considerable variability.

ACKNOWLEDGEMENTS

This research was partially supported by NIH.

REFERENCES

[1] Behrens, S. J. and Blumstein, S. E. (1988a), "Acoustic characteristics of English voiceless fricatives: A descriptive analysis," *J. Phonetics* 16, 295-298.

[2] Shadle, C. H., Moulinier, A., Dobelke, C. U., and Scully, C. (1992), "Ensemble averaging applied to the analysis of fricative consonants," in *Proceedings of the International Conference on Spoken Language Processing*, Vol. 1, (Banff, Alberta, Canada), 53-56.

[3] Xu, Y. and Wilde, L. (1994), "Combining time-averaging and ensemble-averaging in analyzing voiceless fricatives in Mandarin", *J. Acoust. Soc. Am.* 96(5), Pt. 2, 3230. features", *Proc. ICSLP*, 1992, pp. 499-502.

[4] Klatt, D. H. (Chapter 6, unpublished manuscript), "Fricative Consonants."

[5] Wilde, L. F. "Analysis and synthesis of fricative consonants", *Ph.D. Thesis*, Dept. Electrical Engineering and Computer Science, MIT, Feb 1995.

[6] Utman, J. A. and Blumstein, S. E. (1994), "The influence of language on the acoustic properties of phonetic features: A study of the Feature [strident] in Ewe and English." *Phonetica*, 51(4): 221-238.

Cross Language Study of Perception of Dental Fricatives in Japanese and Russian

Seiya Funatsu
Hiroshoma women's University, Hiroshima, Japan

ABSTRACT

The characteristics of Japanese fricatives and Russian fricatives were compared in two ways: (1) by spectrographic analysis and (2) through perception tests of Japanese fricatives by Russian subjects and of Russian fricatives by Japanese subjects. In the case of Japanese, /sj/ is characterized by a higher F2O and lower NF than /s/. In Russian, /s'/ has nearly the same NF as /s/ and is characterized by a higher F2O than /s/. On the other hand, / ʃ / has nearly the same F2O as /s/ and a lower NF than /s/. In Russian subjects, there was a tendency for some of the /sj/ sounds to be identified as /s/. Japanese subjects showed a large confusion between /s'/ and / ʃ /. These results can be explained by the boundaries of native language.

INTRODUCTION

Japanese has two dental fricatives /s/ and /sj/ (which is, phonetically, generally transcribed as [ʃ]) and Russian has three dental fricatives /s/, /s'/ and / ʃ /. When Japanese study Russian and when Russians study Japanese, Japanese fricatives /s/ and /sj/ and Russian fricatives /s/, /s'/ and / ʃ / interfere with each other. This paper describes the results of acoustic analyses of Japanese and Russian fricatives and the results of perception tests.

ACOUSTIC ANALYSIS

Experimental procedure

Japanese speech samples

The speech samples were composed of the following 48 bi syllabic words including nonsense words. The target syllables were word initial syllables.

$$\left\{\begin{matrix}s\\sj\end{matrix}\right\}\left\{\begin{matrix}a\\u\\o\end{matrix}\right\}\left\{\begin{matrix}s\\sj\\p\\t\\k\end{matrix}\right\}\left\{\begin{matrix}a\\i\end{matrix}\right\}$$

In the above list, the words containing the syllables /si/ or /ti/ were excluded because these syllables are phonetically realized as [ʃ i] or [tʃi]. These bi syllabic words were uttered by three male speakers. These 48 speech samples were analyzed.

Russian speech samples

The Russian speech samples for the acoustic analysis were 45 words which have the syllables /sa/, /su/, /so/, /s'a/, /s'u/, /s'o/, / ʃ a/, / ʃ u/, / ʃ o/ in the word-initial position. These words were produced by three male speakers of the Moscow dialect.

Method of analysis

Two acoustic parameters were measured using a spectrograph analysis program on a personal computer. One parameter was the frequency of the peak power in the fricative noise spectrum (NF). Tne other parameter was the onset frequency of the second formant transition of the following vowel (F2O). For measuring the formant, the speech samples were sampled at 10kHz with an accuracy of 12 bits per sample, and FFT analyses were performed. F2O were measured by visual inspection of the spectrogram.

As for noise frequency, the speech samples were sampled at 20kHz and the central parts of the noise periods were extracted using a 51.2-ms Hamming window. NF were determined by visual inspection of the spectrum.

Results

The results of the acoustic analysis are shown in Fig. 1. In the case of Japanese, /sj/ is characterized by a higher NF and lower F2O than /s/. In Russian, /s'/ has nearly the same NF as /s/ and is characterized by a higher F2O than /s/. On the other hand, / ʃ / has nearly the same F2O as /s/ and a lower NF than /s/.

In this figure, it is clear that Japanese /s/ and /sj/ and Russian /s/, /s'/ and / ʃ / all exhibit the coarticulatory effect. Both Japanese and Russian words, when followed by vowel /o/, NF and F2O are lower than vowel /a/. However, in vowel /u/, in Japanese NF and F2O are nearly the same as vowel /a/, while in Russian, NF and F2O are close to vowel /o/. As mentioned above, in vowel /a/, Japanese /sj/ is located between Russian /s'/ and / ʃ / on the NF-F2O plane, but in vowel /o/ and /u/, Japanese /sj/ is located near Russian /s'/.

PERCEPTION TESTS

Perception test of Japanese sounds by Russian subjects

Speech samples

The Japanese speech samples for the perception test were the words which were used in acoustic analysis. These words were sampled at 20kHz and stored in a computer. They were presented to the subjects in random order at intervals of 2s.

Subjects

The subjects were 27 Russian students who had studied Japanese for 1 month in Russia. They were instructed to identify the initial consonant in each word as either /s/ or /sj/.

Results

Table 1 shows the confusion rates between /s/ and /sj/. The over-all error rate is not so large, but there was a tendency that some of the /sj/ sounds to be identified as /s/. But the reason for this type of error is not clear at present and further acoustic analysis of these sounds and perception tests of synthesized sounds are necessary.

Perception test of Russian sounds by Japanese subjects

Speech samples

The Russian speech samples for the perception test were the words which were used in the acoustic analysis. These words were sampled at 20kHz and presented to the subjects in random order at intervals of 2s.

Subjects

The subjects were 38 Japanese students who had studied Russian for 2 months in Japan. The subjects were instructed to identify the initial consonant in each word as either /s/, /s'/ or / ʃ /.

Results

The results are shown in Table 2. It can be seen in the Table 2 that the Japanese subjects showed a large confusion between /s'/ and / ʃ /, but the confusion between /s/ and / ʃ / and the confusion between /s/ and /s'/ were very small. Data in Fig. 1 suggests that on the NF-F2O plane the Japanese phonetic boundary forms an oblique line. The upper left region is /s/, and the lower right region is /sj/. Both of the Russian fricatives, /s'/ and / ʃ /, are located in the region of Japanese /sj/. The above results can be considered as a natural consequence of this acoustic pattern.

Another point to be noted in

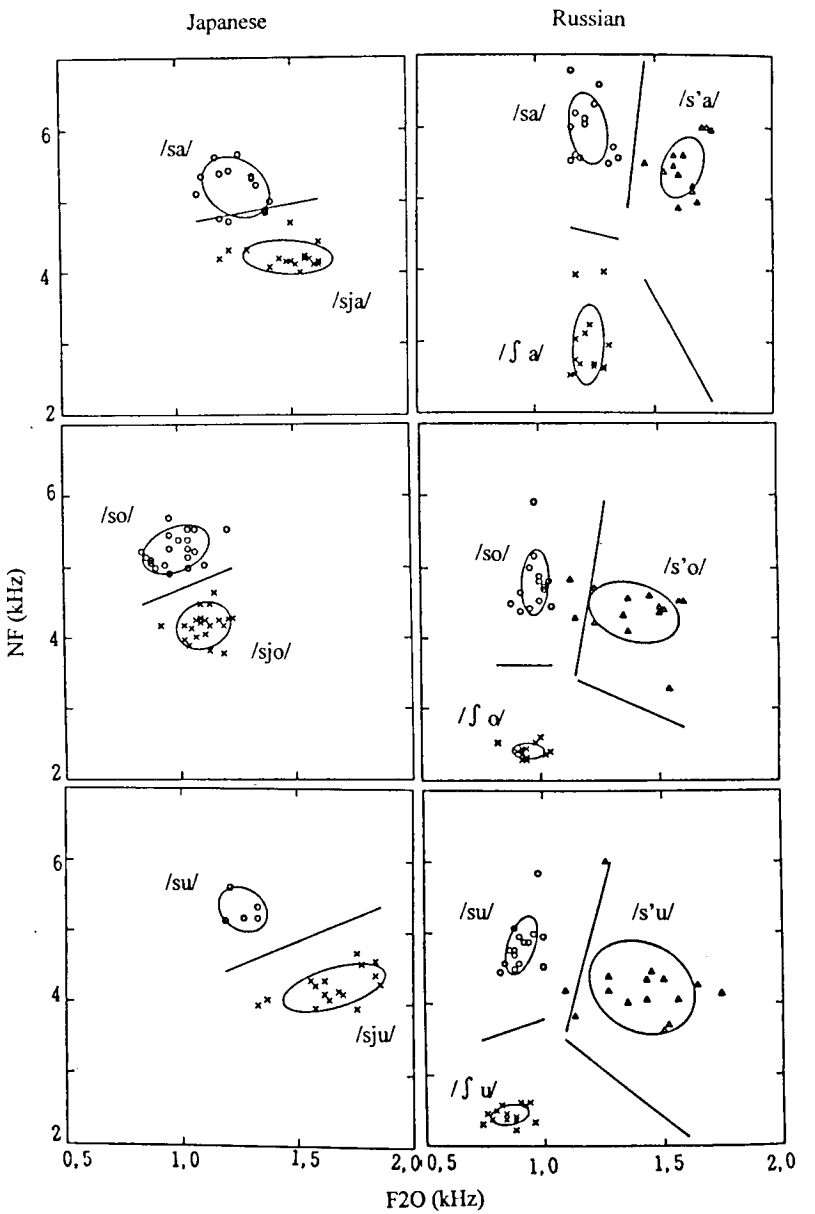


Figure 1. Results of acoustic analyses.
The ellipses represent a 50 percent confidence region on the data values assuming a Gaussian distribution for the data. The straight lines are discriminant functions.

Table 2 is that Japanese subjects identify some instances of Russian /s'/ as /s/. The reason for this type of error is not clear at present and further acoustic analysis of these sounds is necessary.

SUMMARY

As the results of the acoustic analyses show, Japanese /s/ and /sj/ were distinguished mainly by both NF and F2O, while Russian /s/ and /ʃ/, /s/ and /s'/ were distinguished mainly by NF only and F2O only respectively.

In the perception test by Russian subjects, some of /sj/ were identified as /s/, while, by Japanese subjects, they confused some of /s'/ with /ʃ/. These results may explain the effects of native boundaries. But they are not conclusive at present and further experiments are necessary.

Table 1. Confusion matrix of Japanese fricatives by Russian subjects.

following vowel /a/

answer / stimuli	s	sj
s	98	2
sj	6	94

following vowel /o/

answer / stimuli	s	sj
s	99	1
sj	24	76

following vowel /u/

answer / stimuli	s	sj
s	96	4
sj	3	97

Table 2. Confusion matrix of Russian fricatives by Japanese subjects.

following vowel /a/

answer / stimuli	s	s'	ʃ
s	99	1	0
s'	12	69	19
ʃ	1	26	73

following vowel /o/

answer / stimuli	s	s'	ʃ
s	96	4	0
s'	19	51	30
ʃ	2	24	74

following vowel /u/

answer / stimuli	s	s'	ʃ
s	92	6	2
s'	28	29	43
ʃ	3	22	75

VOICE TIMING FOR STOP CLASSIFICATION IN CONVERSATIONAL ENGLISH

Arthur S. Abramson and Leigh Lisker
Haskins Laboratories, New Haven, Connecticut, U.S.A.

ABSTRACT

Acoustic research demonstrating the role of the temporal control of the glottis in separating English "voiced" and "voiceless" stops has mostly used citation forms. We have examined stops in two conversations. For each stop we see whether the voice pulsing is interrupted. If there is a break, we measure the duration of the break with reference to the articulatory release. The results show that temporal control is quite robust even in running speech.

BACKGROUND

Considerable earlier acoustic [e.g., 1], perceptual [e.g., 2] and physiological or articulatory [e.g., 3] work by us and many others [e.g., 4] has demonstrated the significance of the temporal control of the valvular action of the larynx for the division of English stop consonants into the traditional "voiced" and "voiceless" categories. These studies have mostly examined rather deliberate speech: citation forms and short expressions and sentences read aloud.

Although the term "voice onset time" (VOT) has come to be widely used, it was meant by us in the first place to refer to utterance-initial position. Thus, laryngeal pulsing might begin at the moment of closure-release (zero time), before it (voicing lead with time in negative units), or after it (voicing lag in positive units). For widespread varieties of English, with special reference to American English in our work, initial /bdg/ normally have zero-onset of voicing or very short lags of 10 ms or so, although some speakers show voicing lead. Utterance-initial /ptk/ commonly have a rather long voicing lag of some 30 to 40 ms.

The temporal dimension is, of course, not linear in its acoustic manifestations. Voicing lead appears as excitation of the first one or two harmonics during the articulatory closure. Voicing lag appears as noise-excitation of both the release-burst and as much of the formant-pattern as emerges until the onset of pulsing. If the lag is long enough, the turbulence and somewhat attenuated first formant will be heard as aspiration. Experiments with speech synthesis and manipulated natural speech have shown that some several acoustic consequences of voice timing can serve as perceptual cues to the phonological distinction in context-free experiments.

In running speech, with stops occurring immediately after vowels or other consonants, as well as after pauses, the concept of VOT should be broadened to that of "laryngeal timing" or maybe "voice timing" [5]. Tokens of /bdg/ after other voiced consonants or vowels are very likely to have unbroken glottal pulsing in their closures, while /p/ and /k/ before unstressed syllables often have such short voicing breaks as to be heard as unaspirated. (In the latter context "underlying" /t/, as well as /d/, commonly appears in American English as a voiced flap.).

Limiting ourselves for now to the acoustic signal, we wish to assess the stability of the temporal factor in spontaneous fluent English. This is part of our larger interest in the robustness in casual running speech of the differentiating properties and perceptual cues that have long been known for citation forms and careful speech.

PROCEDURE

We recorded about ten minutes of spontaneous conversation held in separate sessions by each of two couples, All four people were native speakers of American English whose minor differences in regional dialects in no way impeded communication. In each couple the man and woman knew each other well and were quite used to talking into microphones; moreover, they were quite at ease with us. Each couple chatted in a relaxed way about personal and professional topics of their own choice without knowing anything of our research goals. Listening to the recordings, we found the conversations intelligible, spontaneous, fluent, and friendly.

After digitizing the recorded speech at 22Kh, we used the Signalyze™ computer program to obtain waveforms and FFT spectrograms. For each conversation, omitting all instances of overlap between speakers and distortions caused by coughing, laughter, and the like, we picked out for analysis all acoustically measurable tokens of the six stops in stressed position, as well as all measurable tokens of /bpgk/ in unstressed position. That is, we excluded the voiced flaps so typical of American English, because any residual contrast in this context between /d/ and /t/ seems to depend upon properties other than voice timing, such as vowel length and quality. We did not include the few instances in our corpus of stops under emphatic stress. We also excluded stops in consonant clusters with /s/ as the first member; here there is clearly no voicing contrast. As for stress, anything not unstressed was taken to be stressed without any attempt at finer gradations.

For each instance of a stop we recorded data on glottal pulsing in the vicinity of the closure and release. With no interruption in pulsing, the item was called "unbroken." An interruption before the release was called a "negative break," and one after the release was a "positive break." The durations of these breaks were measured in the waveforms with reference to the spectrograms. Negative breaks were measured only if there were clear spectral signs of an acoustic discontinuity before the closure with no indication, acoustic or auditory, of a pause. Thus, a stop in utterance-initial position or preceded by a pause could never have a negative break. Also, if a negative break included the closure of a preceding stop, it was not measured. For each stop a "full break" was also entered in our data, whether this was the sum of negative and positive breaks or just the duration of the only one of them available in the utterance.

It is not surprising that in our randomly produced corpus of speech, the stop consonants were unevenly represented across the categories. As a result, we used unpaired two-tailed *t*-tests for assessment of statistical significance.

There were not enough tokens for us to focus on narrower segmental and prosodic contexts. We have postponed any statistical treatment of our two levels of stress.

RESULTS

The means and standard deviations of the full voicing breaks in ms for all four speakers are shown in Figure 1. The average voicing break of the voiceless stops is indeed higher, but there is much overlap of the standard deviations. To this we must add the observation that 84 voiced stops, 62% of that category, had unbroken voicing. They do not appear in the figure.

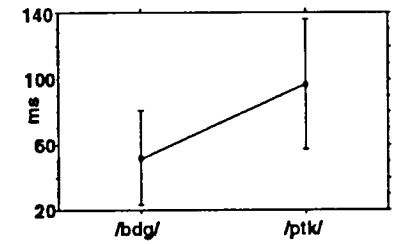


Figure 1. Full voicing breaks: Means and standard deviations for the pooled data of all four speakers. Voiced n= 51; voiceless n = 276.

The data of Figure 1 are broken down into the four speakers in Table 1. Here we see that the difference is significant for all four speakers, although the level is lower for DS and JH. It is interesting to note that the voiceless stops outnumbered the voiced ones by far.

The means and standard deviations of the negative voicing breaks for all four speakers are shown in Figure 2. The data are given separately for the speakers in Table 2, where we can see that while the differences are highly significant for MC and JH, they are not significant for the other two, DS and DL. As for the latter two, however, it must be borne in mind that they do show significant differences in Table 1, so it will be important to see how they fare with positive breaks.

Table 1. Full voicing breaks in ms: Means, standard deviations, and significance levels for the four speakers' unpaired t-tests.

Spkr:	DS	DL	MC	JH
/bdg/				
M	72	62	30	45
SD	27	25	23	18
n	8	22	15	6
/ptk/				
M	106	91	109	82
SD	34	31	49	34
n	41	112	79	44
df	47	132	92	48
t	-2.6	-4.1	-6.1	-2.6
$p <$	.02	.001	.001	.02

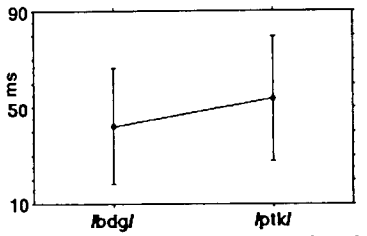


Figure 2. Negative voicing breaks: Means and standard deviations for the pooled data of all four speakers. Voiced n=45; voiceless n=242.

Table 2. Negative voicing breaks in ms for four speakers.

Spkr:	DS	DL	MC	JH
/bdg/				
M	61	48	27	25
SD	15	24	19	19
n	7	22	10	6
/ptk/				
M	51	45	67	54
SD	21	18	33	21
n	40	112	79	44
df	45	132	87	48
t	1.1	-.7	-3.7	-3.2
$p <$	.3, ns	.5, ns	.001	.003

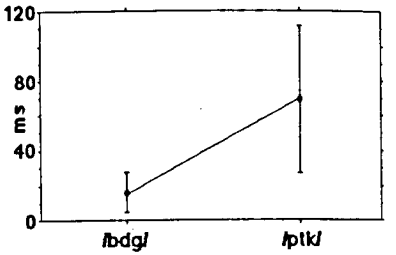


Figure 3. Positive voicing breaks in ms for four speakers. Voiced n=62; voiceless n=293.

Finally, the means and standard deviations of the positive voicing breaks for all four speakers are shown in Figure 3. The data are given separately for the speakers in Table 3. For one of the speakers, JH, the difference is barely significant; for the other three, however, it is highly significant. This also accounts for the significant difference between the full breaks of DS and DL found in Table 1.

Table 3. Positive voicing breaks: Means and standard deviations of the pooled data for the four speakers. Voiced n=62; voiceless n=293.

Spkr:	DS	DL	MC	JH
/bdg/				
M	20	16	12	20
SD	13	12	6	19
n	9	32	15	6
/ptk/				
M	55	45	43	39
SD	19	22	28	23
n	46	125	78	44
df	53	155	91	48
t	-5.4	-7.2	-4.3	-1.9
$p <$	.001	.001	.001	.06,?

DISCUSSION

In the history of speech research certain acoustic properties have been found to have the power to differentiate the phonemes of languages in the production of citation forms or other short utterances. Our research was motivated by a desire to investigate the stability of one of those properties, voice timing, for the distinction between voiced and voiceless stop consonants in American English spontaneous speech.

What with all the contextual redundancy and top-down information present in running speech, one might expect the phonetic rendition of many phonemic distinctions to be somewhat less precise than in more deliberate speech. That is, with so much other information in the discourse, clarity of expression moment by moment ought to be less important. Indeed, just the great temporal variation often observed might blur some distinctions, especially, perhaps, those that include temporal control as an important mechanism.

Despite all the pressures to which such a distinction as consonantal voicing might be vulnerable in running speech, our findings support the general robustness of temporal control of the larynx as an important factor in voicing distinctions in spontaneous conversation. Some generalizations emerge from our sampling of four speakers.

Once the flaps, with their allegedly underlying /d/ and /t/, are eliminated from consideration, it is only the voiced stops that show unbroken pulsing in non-initial position. Thus it is that in our corpus just over 60% of the instances of /bdg/ are distinguished from /ptk/ by this factor alone. As for the rest, relative duration of the voicing break in the region of the closure and release does a rather good job of separating the categories. Even without taking our two levels of stress into account, we find that the voiceless stops have longer voicing breaks than the voiced stops. In addition, it appears that breaks after the articulatory release (positive breaks) bear more of the burden than breaks before the release (negative breaks). Our data are insufficient for examination of narrower phonetic contexts, such as particular vowels.

A preliminary look suggests that a quantitative treatment of the differences linked to stress will remove some of the overlap remaining between the two voicing categories. We plan to do this. Furthermore, we are planning perceptual tests of the validity of our findings.

ACKNOWLEDGMENT

This research was supported by NIH Grant HD-01994 to the Haskins Laboratories. In addition, facilities and help were given to the authors by their respective universities, The University of Connecticut and The University of Pennsylvania.

REFERENCES

[1] Lisker, L. & Abramson, A.S. (1964), "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, pp. 384–422.

[2] Abramson, A.S. & Lisker, L.(1970), "Discriminability along the voicing continuum: Cross-language tests," In *Proceedings of the 6th International Congress of Phonetic Sciences, Prague* (pp. 569–573). Prague: Academia.

[3] Sawashima, M., Abramson, A.S., Cooper, F.S., & Lisker, L. (1970), "Observing laryngeal adjustments during running speech," *Phonetica*, vol. 22, pp. 193–201.

[4] Löfqvist, A. (1980. "Interarticulator programming in stop production," *Journal of Phonetics*, vol. 8, pp. 475–490.

[5] Abramson, A.S. (1977). "Laryngeal timing in consonant distinctions," *Phonetica*, vol. 34, pp. 295–303.

ACOUSTIC AND PERCEPTUAL CHARACTERISTICS OF GEMINATED HINDI STOP CONSONANTS

Nisheeth Shrotriya, A Sada Siva Sarma, Rajesh Verma and S S Agrawal

Speech Technology Group,
Central Electronics Engineering Research Institute Centre,
CSIR Complex, Hill Side Road, New Delhi - 110 012.

ABSTRACT : In the present paper the effect of gemination on the acoustic properties of the stop consonants have been studied. Analysis of duration of silence and preceding vowel shows a strong correlation with the presence of gemination. It is also seen that burst is stronger for geminates and pitch rises abruptly towards the end of the preceding vowel indicating the presence of a geminate. Speech perception tests indicate that about 1.5 times silence duration is required to perceive a geminate.

1. INTRODUCTION: Many researchers have studied acoustic properties of clusters in English and other languages [1, 2, 3]. In the studies of the acoustic properties of certain VCC utterances by means of spectrographic analysis [1] it was shown that the period of the silent interval of a stop consonant varies with the place of articulation and the frequency spectrum of the plosive bursts occur in distinct regions. Repp [2], showed that when the closure period of a naturally produced utterance with two different stop (cluster) consonants in a vocalic context is spliced out only the second stop consonant was heard by the listeners and not the first. To perceive both the consonants (i.e., cluster), 50 to 100 msec. of silence is needed between the two vocalic portions depending on the particular stimuli used. However much longer silent interval (approx. 200 msec) is needed to perceive same stop consonants (i.e. a geminate) [3]. Thus the interval required to perceive a sequence of two intervocalic stop consonants is much longer when the two phonemes are the same as compared to when they differ in place of articulation. To differentiate between the geminates and non-geminates it was decided to undertake the study of acoustic characterstics of Hindi words (natural speech) containing single and double stop consonants.

2. METHOD:

2.1 Speech Material: Sixteen stop consonants which include 8 voiced i.e. /b, d, ɖ, g, b^h , d^h , $ɖ^h$, g^h / and 8 unvoiced i.e. /p, t, ʈ, k, p^h, t^h, $ʈ^h$, k^h/ were used for the present experiment. These were used in between the two vowels of cvCvc and cvCCvc syllables, eg. /sʌtʌr/ and /sʌttʌr/. The preceding and following vowel to the stop consonants was always a short vowel /ʌ/ in our stimuli.

2.2 Data Recording And Analysis: All the words were recorded by five adult male speakers who were native of Hindi and had no articulatory defects, on a TEAC cassette deck (model C-2X) using Senniheiser microphone (Model MD-421). The recorded samples were filtered at 70 Hz - 7 KHz and then digitized at 16 bit, 16 K samples per second using Ariel's DSP 16 card on a PC-AT 386. The speech samples were analysed using a SENSIMETRICS speech analysis package to obtain the audio waveform and digital spectrograms etc. A representative spectrogram of words /sʌtʌr/ and /sʌttʌr/ is shown in fig. 1.

unvoiced stops. In case of the geminates also the same is true. Overall results of the preceding vowel duration shows that it has larger values in the context of non-geminates than that of geminates.

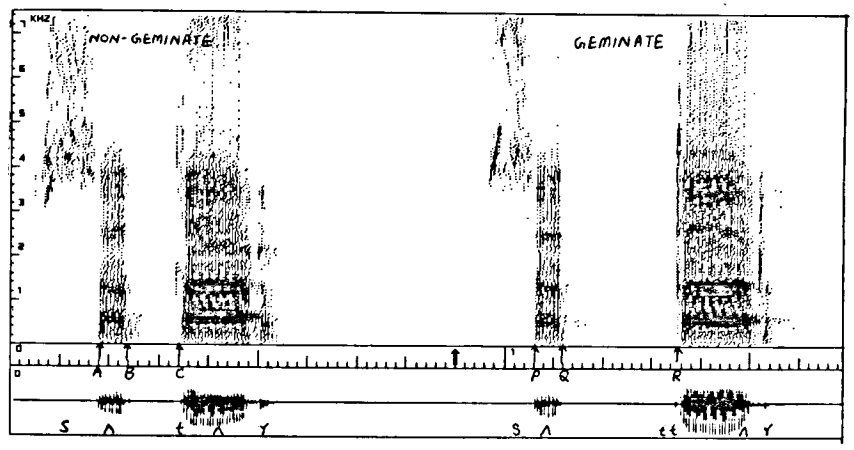


Fig. 1. A representative spectrogram of words /sʌtʌr/ and /sʌttʌr/.
A--->B and P--->Q = Preceding Vowel Duration of the non-geminate and geminate.
B--->C and Q-->R = Closure Duration of the non-geminate and geminate.
C and R are the bursts of the non-geminate and geminate respectively.

3. RESULTS AND DISCUSSION: Table-1 shows the durations of the preceding vowel and closure for the non-cluster (single stop) and their corresponding cluster words (geminates). The table is divided into four categories i.e. UnVoiced-UnAspirated, Voiced-UnAspirated, UnVoiced-Aspirated and Voiced-Aspirated stop consonant.

3.1 Preceding Vowel Duration (PVD): Table-1 indicates that the duration of the preceding vowel was greater for the non-geminate words as compared to its duration for the geminate words. Vowels before voiced consonants have longer durations as compared to those with the

3.2 Closure Duration (CD): Closure duration for the geminates has larger values (about 2.5 times) than that for the non-geminates. There is a significant difference between the closure durations of voiced and corresponding unvoiced stops, whether they are unaspirated or aspirated. The closure duration for both the cases of unvoiced stops i.e. Unvoiced Unaspirated and Unvoiced Aspirated is greater than that of the voiced stops i.e. Voiced Unaspirated and Voiced Aspirated. However in case of the non-geminates the closure duration for Unvoiced Unaspirated and Voiced Unaspirated is found equal to that of the Unvoiced Aspirated and Voiced Aspirated sounds

Table 1: Average durations of the preceding vowel and closure for the non-cluster and cluster words.

	Non-clus. word	PVD (ms)	Avg. (ms)	CD (ms)	Avg. (ms)	Clus. word	PVD (ms)	Avg. (ms)	CD (ms)	Avg. (ms)
U	चपर	70		110		चप्पर	65		230	
V	सतर	65		90		सत्तर	55		225	
U	कटर	70	70	70	90	कट्टर	60	60	230	225
A	चकर	65		90		चक्कर	55		215	
V	कबर	80		85		कब्बर	80		180	
U	गदर	85		70		गद्दर	75		190	
A	कडर	90	85	55	70	कड्डर	85	75	180	185
	लगर	80		70		लग्गर	70		180	
U	सफर	70		105		सफ्फर	60		225	
V	कथर	65		85		कत्थर	50		220	
A	मठर	65	65	90	90	मठ्ठर	60	55	235	225
	पखर	60		85		पख्खर	50		215	
V	बभर	85		75		बब्भर	70		175	
A	लधर	80		70		लद्धर	60		180	
	पढर	85	80	60	70	पढ्ढर	60	65	200	185
	बधर	80		75		बद्धर	70		185	

PVD = Preceding Vowel Duration Avg. = Average
CD = Closure Duration Clus. = Cluster

respectively. It may thus be summarized that in the case of the Unvoiced stops the closure duration is more for both non-clusters as well as geminates.

3.3 Pitch: The changes in the fundamental frequency of the vowel preceding the stop consonant were studied using the CD_LABEL computer software [4]. Fig.2 shows an abrupt rise in pitch towards the end of the preceding vowel indicating the presence of a geminate, whereas it remains nearly constant for non-geminates. This is an important acoustic landmark for the presence of a geminate.

3.4 Burst: Various properties of the burst e.g. duration, spectral shape etc were studied. The results revealed that the burst of geminates is stronger (by about 10 db)

as compared to the burst of non-geminates (Fig. 3). However there are no significant changes in the duration and spectral shape of the non-geminate and geminate sounds.

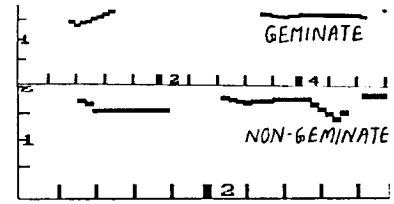


Fig. 2: Comparison of the pitch of the non-geminate and the geminate. [Words /kʌṭʌr/ and /kʌṭṭʌr/].

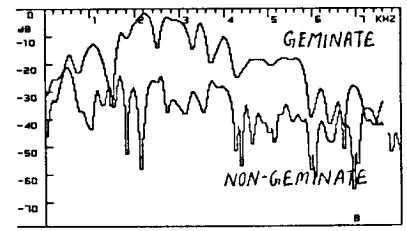


Fig. 3: Comparison of the burst spectra of the geminate and the non-geminate. [Words /pʌḍʰʌr/ and /pʌḍḍʰʌr/]

3.5 Perception Tests: While preparing the stimuli for the perception tests all the non-geminated words were segmented from the centre of the closure and silence was introduced in between these two files in addition to the original silence in steps of 25 msec. These samples were presented to 5 listeners, who were asked to differentiate them as geminate or non-geminate sounds. The results show that a silence duration of 100 msec or less leads to the perception of non-geminate but as soon as the silence duration is made 125 msec. these are perceived as geminate. Thus there is a state of confusion between 100 and 125 msec during which 60% listeners perceive them as geminate while the rest perceive them as non-geminate.

4. CONCLUSIONS: The following conclusions may be drawn from the above studies. (i) For geminates the closure duration is more than double as compared to that for non-geminates. (ii) Closure duration has larger values for UV stops. (iii) preceding vowel duration has greater values for non-geminates while less for the geminates. (iv) Pitch shows an abrupt rise towards the end of the vowel indicating the presence of a geminate. (v) Perception tests show that when the closure durations of non-geminate sounds are increased they are perceived as geminate sounds.

5. ACKNOWLEDGEMENTS: The authors are grateful to Prof. R.N.Biswas, Director, CEERI, Pilani and DoE for encouragement and support. The award of Senior Research Fellowship (SRF) by CSIR to one of the authors (Nisheeth Shrotriya) is highly acknowledged.

6. REFERENCES:

[1] Menon, K.M.N. (1969), "Acoustic Properties of Certain VCC Utterances" Jour. Acous. Soc. Amer, Vol. 46, No.2, pp. 449-457.

[2] Repp, B.H., Liberman, A.M., Eccardt, T. and Pesetskey D. (1978), "Perceptual Integration of Temporal cues for Stop, Fricative and Affricate Manner", Journal of Experimental Psychology : Human Perception and Performance, 441, pp. 621-637.

[3] Pickett, J.M. and Decker, L.R., (1960), "Time Factors in the Perception of a Double Consonant", Language and Speech, 431, pp. 11-17.

[4] P.K. Dhanarajani, et. al.,(1993) "A PC based Graphic Tool for Analysis, Segmentation and labelling of Speech Signals", Proc. 3rd ICAPRDT-93, ISI, Calcutta, India, Dec. 28-31, pp. 326.

THE EFFECT OF VOWEL REDUCTION ON LANDMARK DETECTION

Sharlene A. Liu
Research Lab of Electronics, Dept. Electrical Engineering & Computer Science
MIT, Cambridge, MA 02139, USA
email: liu@lexic.mit.edu

ABSTRACT

In the speech waveform, *landmarks* guide the search for the underlying distinctive features. The landmark detection rate by an automatic algorithm was 94%. An analysis of the prosodic environments in which the landmark detector failed showed that a right-reduced vowel environment caused more misses than other prosodic environments. Consonantal duration was shorter and energy change was smaller in this environment.

1. INTRODUCTION

The proposed model of lexical access uses *landmarks* to guide the search for *distinctive features* [1]. Fig. 1 shows a flow diagram of the lexical access system. In the speech waveform, landmarks are salient points around which important acoustic cues identify the underlying distinctive features. They appear to be perceptual foci, and specify times when certain articulatory targets are to be achieved [2]. After landmarks are detected, distinctive features are extracted in the vicinity of each landmark. The feature specifications associated with each landmark are then organized into a sequence of segments, and the lexicon is accessed by features.

A landmark detection algorithm was developed to automatically locate acoustically-abrupt landmarks [3]. Fig. 2 shows a spectrogram with the acoustically-abrupt landmarks indicated. Acoustically-abrupt landmarks are typically consonantal closures and releases, and other spectral discontinuities caused by velopharyngeal port and vocal fold activity. The algorithm detected most of the desired landmarks, but missed some. In order to understand the circumstances under which it misses landmarks and to improve on the landmark detector, a study of the effect of vowel reduction on landmark detection was conducted. This paper presents the landmark detection algorithm, results of landmark detection, reduced vowel effects, and an acoustic analysis of various reduced vowel environments.

2. LANDMARK DETECTION

This section describes a landmark detection experiment. The database used, the details of the algorithm, and the results will be presented.

2.1 Database

Four speakers (2 female, 2 male) read sentences naturally and clearly. The utterances were recorded with an omnidirectional microphone and digitized at 16 kHz. The signal-to-noise ratio was 30 dB. The acoustically-abrupt landmarks in the utterances were hand-labeled according to the phonetic type of the segments in the vicinity of the landmark (e.g. vowel-stop, vowel-nasal), and whether the landmark designated a closure or release. The reduced vowels (typically /ə/s, syllabic /l/s, syllabic nasals, and sometimes /ɚ/) were also labeled. All other vowels, stressed or otherwise, were considered unreduced.

2.2 Detection Algorithm

The landmark detection algorithm relies on spectral discontinuity and acoustic-phonetic knowledge. It is divided into two stages: general processing and landmark type-specific processing. The output of the algorithm is a series of landmarks specified by time and type.

In general processing, a short-time Fourier transform magnitude (STFTM) is computed and smoothed over 20 ms to remove variations due to glottal pulses and random noise. The spectrum is divided into six bands: 0-0.4, 0.8-1.5, 1.2-2, 2-3.5, 3.5-5, and 5-8 kHz. Band 1 (0-0.4 kHz) keeps track of the turning on and off of voicing. The mid-frequency bands keep track of spectral changes due to sonorant consonantal segments, as well as bursts and the cessation of noise after bursts. In each band, the energy and its derivative are calculated. The peaks in the derivative represent times of abrupt spectral change in a band.

In the landmark type-specific processing stage, the peaks in the derivative direct processing to find three types of landmarks. These three types are: **g**(lottis), which marks the beginning and end of glottal vibration, **s**(onorant), which marks sonorant consonantal closures and releases, and **b**(urst), which designates stop or affricate bursts and points where aspiration or frication ends due to a following stop closure. The **g** landmarks are found from Band 1 peaks. Pairing of landmarks at voicing onset and offset and a minimum syllable requirement are imposed. The **s** landmarks are found from Bands 2-5 peaks during voiced regions delimited by **g** landmarks. A steady-state requirement during the closure of a sonorant consonantal segment and a sufficient high-frequency abruptness are imposed. The **b** landmarks are found from Bands 2-5 peaks during the unvoiced regions delimited by **g** landmarks. A silence period during the closure of a [-continuant] segment, is required.

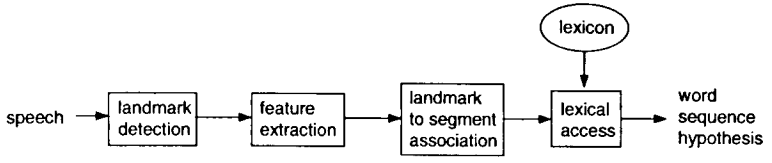


Figure 1: *Flow diagram of the proposed lexical access system.*

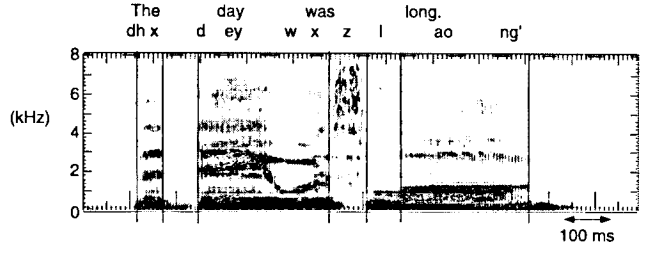


Figure 2: *Spectrogram with acoustically-abrupt landmarks indicated by vertical lines.*

Table 1: *Results of landmark detection.*

# Tokens	1597
Deletion	4%
Substitution	2%
Insertion	10%
Total error	16%

2.3 Results of Detection

Table 1 shows the results of running the landmark detection algorithm on the database. A landmark is considered correctly detected if it is within 30 ms of the hand-labeled landmark and is of the correct type (**g**, **s**, **b**). A *deletion* is a missed landmark. A *substitution* is within 30 ms of the hand-labeled landmark but misidentified by type. An *insertion* is a false landmark. The rates in Table 1 were calculated by dividing by the number of tokens.

From the deletion and substitution rates, one sees that 94% of the landmarks were correctly detected. In terms of phonetic category, almost 100% of the unvoiced obstruents were detected. Voiced obstruents were somewhat more problematic, because voice bars reduce energy abruptness in

Band 1. The algorithm's detection of nasals and [l]s was also lower. One reason is that they are often implemented in a glide-like fashion, so that spectral change is not very abrupt. This is especially true for [l]s. Another reason is that the **s** detector is somewhat context-dependent. Sonorant consonantal segments next to high, back vowels did not always produce a sufficiently large change in energy. The **b** landmarks were detected well for the most part; however, weak bursts and noisy stop closure intervals caused some **b** deletions.

3. PROSODIC EFFECTS

In this section, the effect of vowel reduction on landmark detection in VCV sequences is considered. Table 2 organizes the landmark detection rate by prosodic context. Landmarks occur singularly or in clusters between two vowels. A landmark in *left-reduced* environment means that the preceding vowel is reduced while the succeeding vowel is unreduced. A landmark in *right-reduced* environment is the opposite. *Both reduced* means that both vowels are reduced. *Neither-reduced* means that both vowels are unreduced. The largest error rate occurred for landmarks in right-reduced position, while the smallest error rate occurred for landmarks in left-reduced position. Because the right-reduced environment is the flapping environment for alveolar stops in American English, there is reason to believe this environment causes consonants, in general, to be reduced.

3.1 Constriction duration

An acoustic analysis of the various prosodic environments shows why landmarks in right-reduced environment are harder to detect. One acoustic factor that affects landmark detection is constriction duration. The shorter the duration, the harder the landmark is to detect. The landmark detector relies on detecting energy change. If a constriction is too short, the energy change may be de-emphasized by the smoothing during the 6-band energy calculation. For voiced obstruents, voice bars de-emphasize the energy change in Band 1 even more. The first row in Table 3 shows the average constriction duration of singleton consonants in the four prosodic environments. The constriction duration in right-reduced environment is shortest, explaining in part why

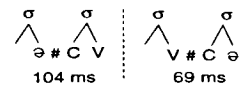


Figure 3: *The effect of vowel reduction and syllable affiliation on consonant duration.*

landmarks in this environment have the lowest detection rate. The constriction duration in left-reduced environment is longest, resulting in a higher detection rate. The durations in both-reduced and neither-reduced environments are in between, in agreement with the detection rates. Of relevance, Turk [4] showed that, within a word, stop consonant constriction durations are shorter in left-stressed environments than in right-stressed environments.

In the above analysis, the syllable affiliation of the consonant was not taken into account. It has been hypothesized that reduction is syllable-affiliated, so that the effect of vowel reduction on a consonant is greater if that consonant belongs to the same syllable as the reduced vowel than if it belongs to a different syllable. To test out this hypothesis, the constriction durations of Table 3 were grouped according to the word affiliation of the consonant. Consonants in word-medial position were not used because of the difficulty of deciding their syllable affiliation. The duration of consonants in left-reduced and right-reduced environments was noted when the consonant was affiliated with the right vowel. Fig. 3 illustrates the two cases considered. Consonants in left-reduced environment had an average duration of 104 ± 30 ms, while in right-reduced environment the average duration was 69 ± 25 ms. The average difference is 35 ms, which is bigger than the 26 ms difference when syllable affiliation was not considered. This finding supports the hypothesis that consonant reduction is affected not only by neighboring vowel reduction, but by the syllable affiliation of the consonant to the neighboring vowels as well.

3.2 Energy change

In addition to constriction duration, the amount of energy change at closure and release also affects landmark detection. The bigger the change, the easier the detection, and vice versa. The

Table 2: *Landmark detection rate, grouped by position with respect to reduced vowels. The number of tokens is given in parentheses.*

	Left-reduced vCV		*Right-reduced VCv*		*Both-reduced vCv*		*Neither-reduced VCV*	
altogether	98%	(444)	87%	(367)	96%	(134)	89%	(390)
+v fric	100%	(41)	81%	(59)	100%	(13)	87%	(77)
+v stop	100%	(52)	80%	(49)	100%	(22)	95%	(59)

Table 3: *Constriction duration and voiced obstruent low-frequency energy change at constriction, grouped by position with respect to reduced vowels. The number of tokens is given in parentheses.*

	Left-reduced vCV	*Right-reduced VCv*	*Both-reduced vCv*	*Neither-reduced VCV*
constriction duration	89 ± 13 ms (151)	63 ± 28 ms (111)	76 ± 21 ms (38)	67 ± 28 ms (106)
+v obstruent energy change	21 ± 5 dB (120)	16 ± 6 dB (144)	18 ± 5 dB (54)	17 ± 6 dB (144)

change in the 20 ms–smoothed, Band 1 energy at closure and release was measured for all voiced obstruents. An energy change at closure was measured by subtracting the lowest energy level (in dB) during the constriction from the energy at a point directly preceding the closure transition in the vowel. At release, the measurement is made with the succeeding vowel. The second row in Table 3 shows that the energy change is 5 dB less, on average, for voiced obstruents in right-reduced environment than in left-reduced environment. The energy changes in the other prosodic environments were in between. This gradation in energy change is consistent with the landmark detector's performance in the four prosodic environments.

4. CONCLUSION

In this paper, the effect of neighboring reduced vowels on landmark detection was studied. Landmark detection is the first step of a proposed lexical access system. It was found that landmarks in right-reduced environment tended to be missed more often than in other prosodic environments, notably the left-reduced environment. An acoustic analysis showed that, in right-reduced environments, the consonantal constriction duration tended to be shorter and the amplitude change smaller than in other prosodic environments. The effect is amplified when syllable affiliation of the consonant to the neighboring vowels is considered.

ACKNOWLEDGEMENTS

I thank Ken Stevens and Stefanie Shattuck-Hufnagel for their guidance. This research was partially supported by NSF.

REFERENCES

[1] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, S. Liu, "Implementation of a model for lexical access based on features", *Proc. ICSLP*, 1992, pp. 499-502.

[2] K. N. Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds", *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, ed. V. Fromkin, Academic Press, New York, 1985, pp. 243-255.

[3] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition", *Ph.D. Thesis*, Dept. Electrical Engineering and Computer Science, MIT, May 1995.

[4] A. Turk, "The American English flapping rule and the effect of stress on stop consonant durations", *Working Papers of the Cornell Phonetics Laboratory*, March 1992, pp. 103-134.

PHONETIC INTERPRETATION OF ACOUSTIC SPEECH SEGMENTS

Knut Kvale

Telenor Research,
N-2007 Kjeller, Norway
E-mail: knut.kvale@tf.telenor.no

ABSTRACT

A crucial problem in automatic speech recognition is the transformation from the continuously varying speech signal to a set of discrete and abstract phonological symbols. The key question is how much phonetic information can be extracted from the speech signal alone without using prosodic, syntactic, semantic or pragmatic knowledge. To address this issue we have analyzed the phonetic content in spectrally homogenous acoustic segments which are selected automatically. Although the segmentation algorithm is language independent and needs no training session, we found that the obtained acoustic segments could be given phonetic interpretations.

1. INTRODUCTION

In the last few years impressive progress has been made in *spoken language systems* (SLS) which make it possible for people to interact with computers using speech. The SLS technology integrates techniques of automatic speech recognition (ASR), natural language processing and human interface facilities. A crucial problem for ASR is the transformation from the continuously varying speech signal to a set of discrete and abstract phonological symbols.

Although listeners tend to perceive speech as discrete sounds following each other in temporal order, the mapping between acoustic events and a linguistic representation is complex, non-linear, irreversible and only partly understood. Thus, the discreteness is not signalled by the stimulus but is imposed on that stimulus by a listener.

The design philosophy of many automatic speech recognition systems has therefore been based on the belief that the acoustic signal does not provide sufficient information to identify the linguistic content of an utterance. Thus, prosodic, syntactic, semantic and pragmatic knowledge has to be utilized to recognize an utterance.

By contrast, experiments in speech spectrogram reading, e.g. [1],[2], have demonstrated that phonemes are accompanied by acoustic features that are recognizable directly from the speech signal without additional knowledge sources. This paper pursues this issue further by analyzing automatically derived stable portions of the speech signal.

2. ACOUSTIC SEGMENTATION

A *segment* is a linear unit anchored in a short stretch of speech by a set of relatively unchanging phonetic feature-values [3]. Thus, *segmentation* can be defined as dividing the speech signal into directly succeeding, non-overlapping stable parts.

Algorithms for automatic *acoustic segmentation* rely on the acoustics only, i.e. they do not assume any phonological information. There are many advantages of acoustic segmentation compared to phonemically based segmentation. Firstly, the speech segments are characterized by acoustic, language independent properties, which can be derived automatically. That is, the calculations are entirely based on signal processing and hence there is no need for explicit modelling or any prior phonological knowledge of the language. Secondly, the automatic subword generation is deterministic in that identical waveforms will be segmented into the same acoustic subword. Thirdly, the acoustic segments often contain highly correlated frames and can hence be quantised, i.e. represented by less data, without losing essential information.

In this paper we analyze the acoustic segmentation calculated by the *Constrained Clustering Vector Quantization* (CCVQ) algorithm [4],[5]. This algorithm recursively computes all possible segment combinations and represents each segment by its centroid, (i.e. its mean spectrum with the present distortion measure). The optimal segment sequence minimises the differences between the spectral frame vectors and the centroid within each segment. That is, the consecutive acoustic segments which yield minimal overall intra-segmental distortions are found. The obtained segments thus exhibit the maximal acoustic homogeneity within their boundaries and the frames within a segment are highly correlated, i.e. steady segments are located.

Phonemically defined units may contain many spectrally homogenous or quasi-stable areas. Thus, acoustic segmentation algorithms may often provide an *oversegmentation* (o.s), i.e. more segments than phonemic labels. As an example, *figure 1* displays a speech waveform and the corresponding broad-band spectrogram which is automatically segmented with the CCVQ-algorithm with 100% o.s., i.e. the number of acoustic segments is forced to be twice the number of phonemes in the utterance. The speech signal in figure 1 is manually segmented and labelled with SAMPA-symbols [6] according to the conventions described in [4],[7].

3. QUALITATIVE EVALUATION

The qualitative analyses of the CCVQ-algorithm were carried out on the Norwegian EUROM0 recording [4],[7]. With 100% oversegmentation typical general trends were (see [4] for details):

- *Plosives* were most often segmented into a closure part and a burst part, such as /k/ in /O:kek/ in figure 1. However, when voiceless plosives succeeded an /s/, as /sp/ and /sk/ in figure 1, the plosive release was weakened and was not marked as a separate segment. If the closure contained some voicing, this was also separated as one segment. Often some alternatives for the beginning of the closure and the end of the burst were given. If the plosive release contained both a burst and an aspiration part, these were marked as two separate segments.
- *Vowels* realised with an amplitude that increased evenly to a maximum value and then decreased towards the next phoneme often contained formant-transitions which were detected by the acoustic segmentation and an acoustic segment boundary was placed near the amplitude top as in the first /i/ in figure 1. (Marking the "centre" of the phonemes is useful for e.g. consistent diphone segmentation for text-to-speech synthesis [8]).
- In the transition from *vowel to silence* the acoustic segmentation algorithm calculated two or three boundaries as for /O:k/ and /ek/ in figure 1. The first one was placed where the intensity reduction began in the higher frequencies, the second (optional) one was placed where almost no intensity was registered in the spectrogram, and the third one was placed where no intensity at all was detected in the spectrogram.
- Segments containing *extralinguistics* (e.g. creaky voice, epenthetic silence, epenthetic sound and lipsmack) were

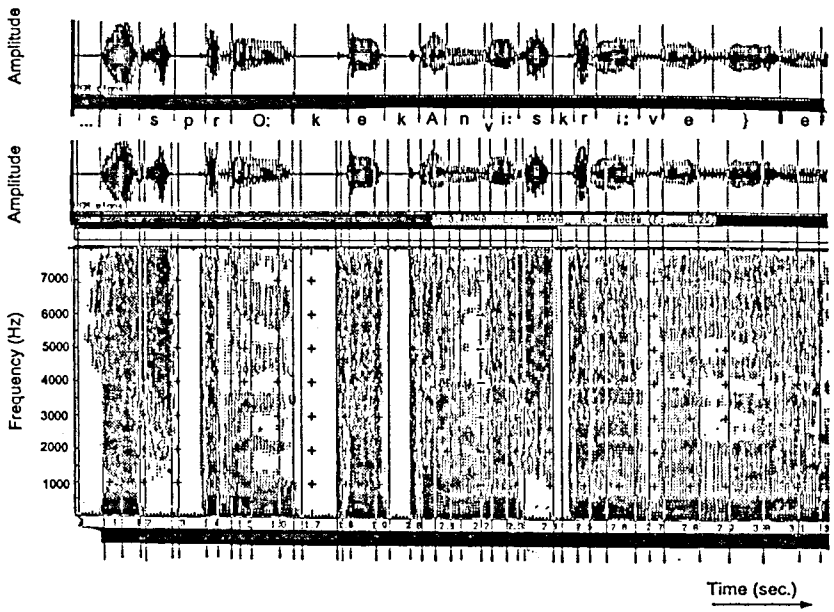


Figure 1 *The sentence "i språket kan vi skrive ue(ndelig)" (=in language we can write inf(initely)) is manually segmented and labelled with SAMPA symbols [6] shown under the waveform in the top row of the figure. In the waveform below and in the broad band spectrogram the acoustic segmentation boundaries with 100% over-segmentation are shown. After [4].*

marked off separately. In figure 1 we notice that the creaky voice area between /e/ and /}/ is one acoustic segment.

• When the apico alveolar tap realisation of /ɾ/ showed up in the spectrogram with an extra voiced sound with formant structure [9], i.e. an epenthetic schwa as in /sprO:/ and /skri:v/ in figure 1, the schwa and [r]-closure were segmented into separate acoustic segments.

4. OVERSEGMENTATION

Obviously it is preferable to keep the oversegmentation factor (o.s.-factor) as low as possible while still achieving high coincidence with manual segmentation. This section summarizes the performance of CCVQ-segmentation as a function of oversegmentation:

• Boundaries computed with a lower o.s.-factor remained fixed when increasing the o.s.-factor. That is, the effect of increasing the o.s.-factor was to split the segment(s) with highest intra-segmental distortion. Actually, spectrally stable segments were not divided even with 200% oversegmentation.

• The CCVQ-algorithm searched for stable segments, and the boundaries were placed in transient areas because vectors from these areas increase the intra-segmental distortion. As the o.s.-factor increased, transition areas could be segmented into several short acoustic segments, providing several alternative boundaries. This reflects the segmentation problem of placing a boundary between two sounds at one, single, "correct" time instant.

• With more than 75% o.s., the acoustic segmentation obtained high coincidence with the corresponding manually placed boundaries. The few deviations from manual segmentation were mainly due to:

i) Some *half-way*, *mid-point*, or *symmetric* conventions used in the manual segmentation, e.g. the convention in [4] of placing the boundary in the middle of the creaky voice area between two vowels (instead of at the end of the segment where abrupt changes often occur). If this area was spectrally stable, the acoustic segmentation assigned boundaries at the ends of it.

ii) "*Impossible cases*", where no boundary cue was seen in the waveform or spectrogram, and the human labeller has placed the boundary rather arbitrarily or based the decision on listening only.

iii) "*Squeezed in segments*", i.e. a phoneme which is perceived when listening to it in context but which is without any corresponding visible acoustic cues in the waveform or spectrogram, was often squeezed in as a very short segment between the phonemes with clear acoustic cues, e.g. as /v/ in /nvi:/ in figure 1.

5. CONCLUSIONS

The CCVQ-algorithm isolated spectrally stable portions of the speech signal. The stable segments were not divided even with a high degree of oversegmentation.

When the number of acoustic segments was forced to be twice the number of phonemes in an utterance, most of the acoustic segments obtained by the CCVQ-algorithm could be given a phonetic interpretation. In addition, quantitative analyses in [4] have showed that the acoustic segment boundaries coincided equally well with the corresponding manual segmentation for English, Danish, Norwegian and Italian (manually annotated by native phoneticians).

Since the acoustic segmentation algorithm is capable of isolating identifiable sub-phonemic segments consistently, it can be useful for speech analysis and automatic speech recognition based on acoustic subwords. The CCVQ-algorithm may also be used as a language independent pre-segmenter tool for manual segmentation of e.g. diphones for text-to-speech synthesis. When this tool is accompanied by conventions for which boundaries to select for the various phoneme transitions, it will reduce the randomness in manual segmentation.

REFERENCES

[1] Zue, V.W. and Cole, R.A. (1979), "Experiments on spectrogram reading", Proc. International Conference on Acoustics, Speech and Signal Processing, pp. 116-119.

[2] Zue, V.W. (1989), *Speech Spectrogram Reading - An Acoustic Study of English Words and Sentences*, Course at University of Edinburgh.

[3] Laver, J. (1994), *Principles of phonetics*, Cambridge University Press.

[4] Kvale, K. (1993), *Segmentation and Labelling of Speech*, Doctoral thesis, Norwegian Institute of Technology.

[5] Svendsen, T. and Soong, F.K. (1987), "On the Automatic Segmentation of Speech", Proc. International Conference on Acoustics, Speech and Signal Processing, pp. 3.4.1-4.

[6] Wells, J.C., et al (1992), "Standard Computer-Compatible Transcription", in *ESPRIT PROJECT 2589 (SAM): Final Report; Year Three; 1.3.91-28.2.92*, SAM-UCL-037.

[7] Kvale, K. and Foldvik, A.K. (1991), "Manual Segmentation and Labelling of Continuous Speech", ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication", pp. 37.1-5.

[8] Kvale, K. (1995), *Manual segmentation of logatomes for diphone-based text-to-speech synthesis*, Scientific Report 2/95, Telenor Research.

[9] Kvale, K. and Foldvik, A.K. (1992), "The multifarious r-sound", Proc. International Conference on Spoken Language Processing, pp. 1259-1262.

STRONG CUES FOR IDENTIFYING WELL-REALIZED PHONETIC FEATURES

A. Bonneau, S. Coste-Marquis and Y. Laprie
CRIN-CNRS & INRIA-Lorraine, Nancy, France

ABSTRACT

We show that there exist cues, called strong cues, which allow some phonetic features to be identified or eliminated with certainty. The use of strong cues brings numerous advantages in ASR, including reliability, and reduction of the search space. With a view to validating our approach, we have defined a first set of strong cues characterizing the place of articulation of French stops. The firing rates of these cues are relatively high. The definition of strong cues can be extended to other features and is useful for different ASR approaches.

1 INTRODUCTION

Recognition techniques generally rely on the use of continuous criteria (as, for example probability density functions for statistical methods). A major disadvantage of this kind of decision is that exemplary realizations of some cues cannot be taken into account and thus cannot lead to definite decisions. Hence, we propose to define strong cues which allow some phonetic features to be identified or eliminated with certainty. Strong cues are both discriminating and well pronounced (according to its realization, a cue can be strong or weak). They must not be confused with main or robust cues.

2 ACOUSTIC CUES

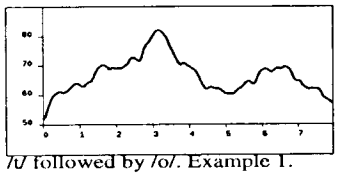
We choose to use context-dependent acoustic cues. For this purpose, we distinguish three classes of vowels: high front vowels (called from now on front vowels), open front and central vowels (called central vowels) and back vowels.

2.1 Description

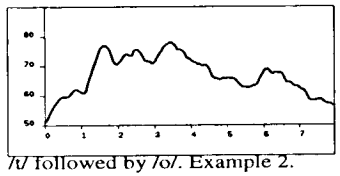
Cues provided by the burst

We use a context-dependent compactness cue to distinguish the velars which are compact from the labials and the dentals which are diffuse. A burst is considered compact if the energy is concentrated around a particular peak. This peak must be the most prominent peak of the spectrum and must be located in a restricted frequency region corresponding to the (context-dependent) region of the spectral maximum of velars. The compactness cue may be more or less pronounced according to the pronunciation of the consonant. We mean that the level of energy concentration may be more or less high and the frequency of the peak more or less close to the expected value. When the cue is very pronounced, the velar place can be identified with certainty.

We use a context-dependent acuteness cue to distinguish dentals from labials. High frequency peaks (above 2200 or 2500 Hz) generally dominate the dental spectrum while low frequency peaks often dominate the labial spectrum. This spectral configuration may not be so clear (or even may not be observed) for all the pronunciations of a consonant. For example, with regard to dental spectra, we know the existence of a peak near the F2 dental locus [1]. This peak sometimes dominates the spectrum. An other example, labials preceding front vowels are often more acute than what is reported in the literature and we frequently observed labial spectral maxima at 2500 or 3000 Hz, which may create confusions with dental consonants. Nevertheless, we believe that, when the dental cue is particularly pronounced, the identity of the consonant is not to be questioned. Figure 1 shows the spectra of two dental consonants, each followed by the back vowel /o/. The two consonants were uttered by the same speaker, in the same sentence, and were both in stressed position. Our stop recognition module, described in [2], detected a strong dental cue for the first example -i.e. in this vocalic context, the identification of /t/ was certain-, while a weak dental cue was detected for the second example -/t/ was the preferred but not the only solution-.



/t/ followed by /o/. Example 1.



/t/ followed by /o/. Example 2.

Cues provided by the transitions

Below are only described CV syllable transitions. The opposite transition trajectories are used for the corresponding VC syllables.

/p,b/. Labialisation lowers F2 and F3 of unrounded vowels. Thus transitions between labials and subsequent unrounded vowels are rising. For rounded vowels, transitions are relatively flat. Since the labial articulation does not require the active participation of the tongue, vowel-to-vowel movement involving this articulator may happen during the articulation of the consonant. As a consequence, if one vowel of a VCV sequence is far less anterior than the other, the expected transition trajectory may not be observed for this vowel.

/t,d/. F2 transitions between vowels and dentals come from a well determined frequency area: the F2 dental locus (around 1500-2000 Hz). Consequently, CV transitions are falling before back vowels (in this context, they constitute a very clear and systematic cue), are falling or flat before central vowels and are flat or rising before anterior vowels.

/k,g/. Cues vary according to the vocalic context. When the vowel is a back vowel, transitions are relatively flat. When the following vowel is a central or a front vowel, F2 and F3 move away (velar pinch). In cases of great coarticulation, F2 and F3 of central vowels, normally spaced, are close together even at the vowel center. Consequently, the velar pinch is not observed.

2.2 Strong cues

As previously observed, a cue can be more or less pronounced according to the realization of a given feature. When a discriminating cue is well pronounced, the feature identification is not only certain but also direct since the detection of other cues becomes useless. We call such a cue a strong cue. Moreover, some spectral configurations are never observed for a given feature. Such a knowledge is valuable for ASR since it allows to eliminate a feature with certainty. We thus propose to define strong preference cues as well as strong exclusion cues. These cues will be used as "anchor points" in the acoustic-phonetic decoding step (see section 4).

Strong cues provided by the burst

Because of the flat spectrum and of the great spectral variability of labials, we do not believe we can define effective strong cues for these consonants.

Context /u, o, ɔ/

Strong preference cue for velars: the energy is concentrated around a prominent peak situated in front of the F2 of the subsequent vowel.

Strong exclusion cue for velars: the lack of a peak in the vicinity of the F2 of the following vowel.

Strong preference cue for dentals: a relatively prominent peak in high frequencies (between 2200 Hz and 4000 Hz).

Context /a, ɛ, œ/

Strong preference cue for velars: the energy is concentrated around a promi-

nent peak situated in the region delimited by the F2 and the F3 of the following vowel.

Strong preference cue for dentals: a relatively diffuse prominent peak in high frequencies (between 2400 and 5000 Hz).

Strong cues provided by the transitions

Context specification allows us to define strong preference cues. For example, a rising CV transition is a discriminating cue for labials followed by a central vowel. Nevertheless, we do not want to define strong preference cues for transitions. The main reason which motivated our choice is that formant tracking is a particularly difficult task at a boundary between vowel and consonant. Nevertheless, it is easier to establish that a trajectory do not correspond to the expected one. If the observed trajectory is drastically opposite to it, the definition of strong exclusion cues is generally possible. Note that strong cues are only considered if the quality of the corresponding acoustic detectors (formant tracking algorithm) is good.

Strong preference cue for velars followed by a central vowel: F2 and F3 come close together.

Strong exclusion cue for labials: F2 of one vowel (in a V-stop-V sequence) takes the opposite direction to that expected, although the other vowel is not more anterior than it.

Strong exclusion cue for dentals: F2 certainly does not come from the dental locus.

3 EXPERIMENTAL RESULTS

3.1 Methodology

We tested the strong cues described on three French corpora. The first corpus (extracted from BDSONS) we used contains isolated words spoken by 5 male speakers. The two other corpora are made of continuous speech. VERLOC was recorded in an office and is constituted of 17 sentences spoken by 16 male speakers (3, 4 or 5 repetitions). The last one contains 22 read sentences made up of stops and vowels, each sentence is pronounced 3 times by 4 male speakers.

The tested items /? stop V/ were extracted from the corpora and hand-labelled. The burst analysis and formant tracking algorithm come from Snorri [3].

3.2 Results

We give the firing rate of strong cues on Table 1 (preference and exclusion cues). We used 758 unvoiced stops and 1769 voiced stops for central vowel context and 610 unvoiced and 933 voiced stops for back vowel context.

	Excl. /p,b/	Excl. /t,d/	Excl. /k,g/	Pref. /t,d/	Pref. /k,g/
p	–	12.5	35.5	–	–
t	1.5	–	41	46.5	–
k	≈ 0.5	18.5	–	–	48.5
b	–	17	44	–	–
d	6.5	–	36.5	23.5	–
g	0.5	27.5	–	–	17.5

Table 1: Firing rates of strong cues for back vowel context (%).

The fact that strong cues alone allow a direct conclusion in more than 40% of cases in the back vowel context validate our approach, even if the back vowel context is undoubtely the most favourable context. Partial results obtained for the central vowel context show that the exclusion cues are only slightly less discriminating than for back vowels (5% lower on average) and thus remain very interesting. However it appears that the burst decomposition algorithm designed for back vowels should be adapted to other vowel contexts in order firing rates of strong cues defined on the burst become higher.

4 ADVANTAGES OF USING STRONG CUES

With regard to the automatic speech recognition, the use of strong cues brings numerous advantages such as the reliability of the information provided, the reduction of the search space, and the possibility to maintain the coherence during the decoding process. Let us develop the two last points.

4.1 Reduction of the search space

If at least one strong cue is detected, the number of acoustic cues and the number of phonetic and lexical solutions are reduced. Indeed, the search for weak cues becomes useless when a preference cue is detected, and limited when an exclusion cue is detected. Taking the decision is then simpler since there is no need to use score combinations, always difficult to turn out. The number of phonetic solutions is reduced: only one solution is proposed when a preference cue is observed, one or several solutions are definitely dismissed when an exclusion cue is detected. Furthermore, using strong cues as confidence islands decreases the search space of the lexical module [4]. For example, if, at the beginning of a word, the dental feature is identified (or dismissed) by a strong cue, the word proposed as the solution must (or must not) begin with a dental consonant.

4.2 Consistency of the decoding process

Strong cues can be used to maintain the consistency of the reasoning during the decoding stage. This strategy has been adopted by the system Daphné [4]. The principle is that strong cues must not contradict one another. Then, in case of conflict between two strong cues, the context in which these cues have been detected has to be questioned. This context includes essentially the segmentation stage, the acoustic detectors (the formant tracker, the burst detector...), and the scope of the coarticulation phenomenon. Let us give a real example we have encountered with the system Daphné. In a VstopV context, the following two strong cues were detected: an exclusion cue for dental stop provided by the transition situated on the left-hand side of the consonant, a strong preference cue for dental stop provided by the burst. This contradiction had to be explained. The system questioned the context and found that the stop closure was relatively long. A new hypothesis, the presence of a stop cluster, removed the contradiction and led to the right solution.

5 CONCLUDING REMARKS

We have shown that, thanks to the use of strong cues, the stop place of articulation can be identified or eliminated with certainty in numerous cases. Strong cues can also be defined for other features, particularly for the place of articulation of fricatives and for the features characterising the manner of articulation. We also believe that the use of strong cues, which brings numerous advantages for knowledge based recognition systems, can be of great interest for systems based on statistical methods.

REFERENCES

1 S. Blumstein and K. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, 66:1001–1017, 1979.

2 A. Bonneau, S. Coste, L. Djezzar, and Y. Laprie. Two levels acoustic cues for consistent stop identification. In *Proc. of the International Conference on Spoken Language Processing*, volume 1, pages 511–514, Banff (Alberta, Canada), 1992.

3 Y. Laprie and L. Mercier. Un environnement logiciel pour un atelier phonétique. In *Actes des XXèmes Journées d'Étude sur la Parole*, pages 209–214, Trégastel, 1994.

4 S. Coste-Marquis. Interaction between most reliable acoustic cues and lexical analysis. In *Proc. of the International Conference on Spoken Language Processing*, pages 1187–1190, Yokohama (Japan), 1994.

ACOUSTIC CUES ON INTERSYLLABIC BOUNDARY IN (C)VNNV STRUCTURE

Xiaoxia Chen
Institute of Linguistics, CASS, Beijing, China

ABSTRACT

The purpose of this study is to investigate the acoustic cues on inter-syllabic boundary in the (C)VNNV structure of Standard Chinese by examining the variation on amplitude, formant frequency and duration during the period of the -NN-. Relevant acoustic measurements were made and a limited perception test was conducted. The experimental results show that some acoustic cues do exist in the boundary between two syllables. That is, an amplitude valley occurs between -N and N-, the nasal duration of -NN- is systematically longer than single -N or N-, and the formant pattern of -NN- is different from that of -N or N-. Moreover, according to the perception test, the transitional area and its duration between two syllables can be estimated, and it roughly matches with the location of the amplitude valley.

I.INTRODUCTION

This report is a part of the research work on intersyllabic juncture in Standard Chinese. Juncture, as the cue of segmental boundary, is a common phenomenon between different speech units. It is true that the juncture is quite complex and rather difficult to deal with. However, it is an important issue from both theoretical and practical point of view, so, it cannot be avoided in speech research.

With respect to Standard Chinese, Hockett(1947) and Chao(1968) mentioned some phenomena in connection with juncture, but experimental approach was first made by Xu(1986). He found that the most ambiguous situation occurred between two adjacent [n] nasals. Since the nasal [n] is a consonant, it can serve both as initial and final ending in a syllable in Standard Chinese. Consequently, it often gives rise to some difficulty in determining the syllabic boundary between the adjacent syllables in (C)VNNV. He only pointed out that the duration of -NN- in (C)VNNV structure is longer than that of single -N or N- in monosyllables. We try to carry this research forward to see if there are any regular patterns that can be regarded as the further acoustic cues of intersyllabic boundary during the period of -NN-.

II.MATERIALS AND METHODS

In this study, 24 (C)VNNV disyllables were selected as the experimental words. They are all normally stressed and meaningful words. These materials were produced by two native male speakers, and the audiorecording was made with P-Kenwood Recorder in the soundproof of the Phonetic Lab, Institute of Linguistics, CASS. All the materials were stored as waveform in the disk. Acoustic analysis and measurements were made through Kay 5500 and ASL program.

In normal spectrogram, the amplitude contour of -NN- period appears rather flat and difficult to differentiate any distinctive variation. In order to further observe its precise structure, the amplitude contours of all the -NN- parts were extended by using ASL program, and the measurements were made at the point of every ten milliseconds.

In order to examine time-varing situation of -NN-, formant frequencies under 3KHz were measured at the onset, midpoint and offset of the -NN- respectively according to their spectrogram.

The durations of the nasals both in monosyllables and disyllables were got from the spectrogram on 5500 sonagraph, and the durational ratios of -N or N- to the whole words were calculated as well.

Another durational measurements related to the transition portion between -N or N- in the -NN- were made by an editing and listening test through 5500 sonagraph. That is, first, the duration of the nasal ending -N from the onset of the -NN- forward was increased until the first syllable heard well; then gradually the duration of the initial N- from offset of the -NN- backward was increased until the second syllable was heard well. The rest of -NN- was the transition between the final ending -N and the initial N-, and its duration could be calculated. Thus, an intersyllabic boundary is supposed to falling into this period.

III.RESULT AND DISCUSSION

Experimental results can be summarized by Fig.1 to fig.3 and Table1.

3.1 Amplitude

Fig.1 shows two examples of the extended amplitude contour of the -NN- period from the data measured in this investigation. It is clear that there is an amplitude valley during the -NN- period. It indicates that there must exist a turning area of the energy, though the turning point cannot be determined precisely. Usually, the energy of a syllable is gradually rising at the beginning and falling at the end, so the turning area observed here should be regarded as a cue of the intersyllabic boundary.

Table 1. The average duration ratios of initial N-, final -N in monosyllabeles and of -NN- in disyllabic words, as well as the transition durations between -N and N- in -NN-.

		N-	-N	-NN-	Tn
C	rate%	14.5	17	27	10
	sd	5.8	6.5	6.3	3.6
	n	40	20	24	24
M	rate%	22	23	32	14.5
	sd	3.7	5.6	8.0	5.5
	n	44	49	23	23

3.2 Duration

Duration data are given in Table1, where the ratios are calculated from the mean durations to -N or N- to entire monosyllable and

of -NN- to the whole disyllabic word respectively. According to the data shown in this table, the -NN- is evidently longer than either the single -N or single N-. This further confirms Xu's report(1986), and it is also similar to that in Tamil (Balasubramanian, T, 1982). It means that there may be a boundary existed somewhere during the -NN-.

The data in the last column of Table 1 are the duration of transition between -N and N- in the -NN- determined by editing and listening test. It makes the boundary location in a narrower and more limited period, which period can be seen from the spectrogram shown in Fig.1.

Moreover, an extended amplitude contour of the -NN- is given in Fig.2 as well. As compared with the position of the transition and the corresponding amplitude contour, an interesting phenomenon can be observed, that the amplitude valley roughly matches with the period of the transition. It further indicates that the intersyllabic boundary does locate in this period.

3.3 Formant pattern

Fig.3 shows the formant patterns of -N in syllable AN, N- in syllable NA and -NN- in word ANNA. Which are drawn according to the frequencies measured from this investigation. From the comparison of these patterns, we can see that the entire pattern of -NN- is evidently different from that of -N or N-. The formant frequency of -NN- varies from the onset to the offset, roughly speaking, the former part is more similar to that of -N, and the later to that of N-, and the middle part is different from either of -N and of N-.

CONCLUSION

Experimental results described above indicate that there do exist some acoustic cues for the intersyllablic boundary: a dinstinctive valley of amplitude occurs in the period of the -NN-; the differences both of duration and of formant pattern are found between the -NN- and the -N or N-; an estimated transition period is determined and the amplitude valley falls into this period.

ACKNOWLEDGMENT

I am grateful to Prof.Cao Jianfen for her help during writing the report.

REFERENCES

Balasubramanian,T.(1982), Intervocalic double nasal and lateral consonant articulations in Tamil, *Journal of Phonetics*, Vol.10, No.1.

Chao,Yuan-ren(1968), *A Grammar of Spoken Chinese*, University of California Press, Berkley, Los Angeles, London.

Hockett,C.A.(1947), Peiping Phonology, *Journal of the America Oriental Society*, 67.

Lin, Maocan & Yan, Jingzhu (1991), Coarticulations in the Zero-Initial Syllables And Its Acoustic Characteristics in Standard Chinese, *RPR-IL(CASS)*.

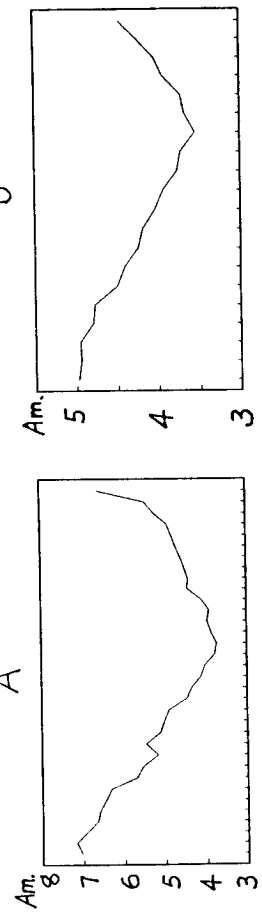


Fig.2 The diagrams of amplitude tendency in -NN- portion,A is subject M(22),B is subject C(24)

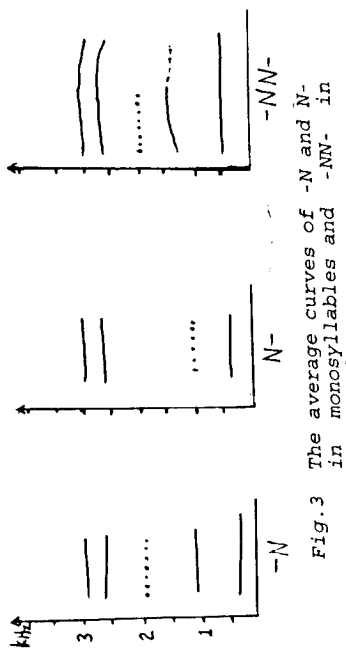


Fig.3 The average curves of -N and N- in monosyllables and -NN- in disyllables

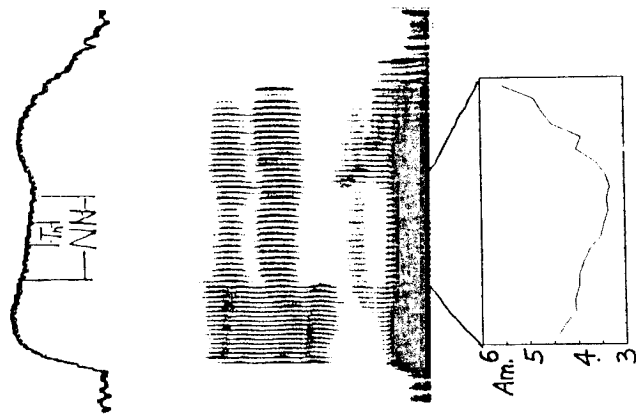


Fig.1 An example of measuring the -NN- in disyllable, the word is Yunnu(yn nu)

MODELING LIP CONSTRICTION ANTICIPATORY BEHAVIOUR FOR ROUNDING IN FRENCH WITH THE MEM (*MOVEMENT EXPANSION MODEL*)

C. Abry and T.M. Lallouache
ICP URA CNRS n° 368 INPG/ENSERG Université Stendhal
BP 25 F-38 040 Grenoble Cedex 9

ABSTRACT

A new model called MEM, *Movement Expansion Model*, is proposed as an alternative to current anticipatory models. Initially developed to deal with one of the correlate of vocal-tract lenghtening (upper lip protrusion), this model is presently extended to the other main component of rounding, the modulation of between-lips area, which time course has never been integrated in the frame of anticipatory models, in spite of its crucial role in acoustics.

1. INTRODUCTION: Protrusion MEM

We are currently developing a *Movement Expansion Model* (MEM [1]), as an alternative to other models available in the field of speech anticipatory behaviour: the so called *look-ahead* [LA], *time-locked* [TL] (now *frame* or *coproduction*) and *hybrid* ("LA+TL") models [2].

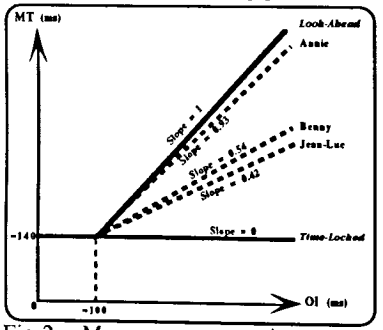


Fig.2– Movement expansion model (MEM): dotted speaker specific regression lines and LA & TL predictions (from [1], see text).

In this model – which classically dealt specifically with upper lip protrusion – in French [i(CCCCC)y] transitions [1], movement time MT was shown to be dependent on the effective duration of the string of consonants (obstruence interval *OI*) produced in the transition from vowel to vowel. We showed that this movement expansion was linearly related to the duration of *OI*, the slope of this relation being speaker-specific. MEM specified for each subject a basic duration for protrusion *MT*, typically 140 ms for [iy] and [iCy] (Fig. 1), as well as an expansion function, starting from about 100 ms *OI* for [iCy]. Speaker-specific parameterization is clearly not in favour of a generalization of either the LA nor TL models or current modified versions. For short, in our [i] to [y] transitions, the anticipation of the protrusion movement is not determined by the end of the unrounded vowel [i], like in LA: in Fig. 1, only one subject, Annie, displays such a behaviour, with a slope near 1. Neither is anticipation determined in a fixed way in relation to the acoustic onset of the rounded vowel [y], like TL: no subject had a relatively constant *MT*, i.e. zero slope was not observed for our other two speakers, Jean-Luc and Benny, whose coefficients are about 0.5.

2. NEED FOR EXTENSION: Constriction MEM

The purpose of the present study was to extend this result to the description of the time course of between-lips area. This area parameter is known to be the most responsible for acoustic changes [3]. It has, to our knowlegde, never been integrated into anticipatory models and other students in the field have called for it [2]. Since we have the image processing system to measure accurately this parameter [4], it was planned for us since the beginning of our work that we would have to use it within an anticipatory model. We were only refrained to do so by the inspection of these lip-area temporal functions. These are rather "bumpy" due to the action of the jaw recruited to produce coronal consonants, like [s,t,l]: the elevation of this carrier articulator diminishes the area, without any active movement of the lips.

In addition, in [1] we could use only 3 out of the 4 subjects we recorded initially, since one of them (in spite of being French !) displayed quite no upper lip protrusion, except a few 10th of millimeter, compared to a range of about 8 mm for the others.
We will show that when we use main events to describe such area profiles, it is possible to predict the time course of the constriction of the vocal-tract output with the same MEM model we used for protrusion.

3. METHOD

Processed face signals (27,000 frames) were the same as in [1], with transitions ranging from [iy] to [ikstsky]. Labelling of audio and video signals was also the same for *OI* interval and kinematic events.
The procedure used to detect events on temporal functions of between-lips area for [i] to [y] transitions was specially designed to maximally avoid small consonantal perturbations (hence ambiguities). Since the obtained curves reflect pretty accurate measurements, it was not chosen to smoothe systematically such perturbations and the weight used for cubic splines to get a continuous function was high enough. So kinematic events used for protrusion and obtained from derivatives were discarded, being too sensitive. We finally characterized these movement profiles with 5 events. First, we considered that when a 10% value of Area Amplitude (10% [Max.Area-Min.Area]), was reached (*10%Area.On*[set]), held or diminished, then increased (*10%Area.Off*[set]), we could safely determine a "Hold" (*H*) phase where acoustic efficiency of constriction was ascertained enough. It follows that we detect *90%Area.On*, reflecting the onset of the constriction movement towards [y], and of course *Max.Area* and *Min.Area*. The "Time Falling" (*Tf*) phase begins with *90%Area.On*, and ends with *10%Area.On*. We will finally, among other phase combinations, use *Tf+H* as a global phase to get the best overall prediction of movement expansion.

4. RESULTS

4.1. Constriction phases and *OI*

Taking advantage of the procedure we used for the study of upper lip protrusion, we chose not to begin by the examination of articulatory events referenced to the acoustic domain, but we searched for correlations between the two flows, i.e. articulatory phases with OI, without a common reference event (in order to avoid part-whole correlation artefact [5]).
Correlation coefficients ($r = 0.32$ at $p = 0.01$) were calculated for all phases without [iy] and with [iy] in order to test intercept values, since the test of the MEM (contrary to other procedures used currently by other students [2]), allows to evaluate, from all samples where *OI* is different from zero, the prediction of the basic simple transition gesture duration.
Figs. 2a-d show the piecewise fitting for each speaker in *Tf+H* and *OI*. Notice that there is no real temporal continuum between one-consonant sequence and the others. Other prosodic factors should certainly be manipulated (for example rate) to cover the whole range of variation of this obstruence interval (*OI*), variously filled, depending on the habits of each speaker. r values corresponding to calculations without [iy] are all very high (from 0.87 to 0.99), contrasting sharply, like for upper lip anticipation, with quantitative data published for English [2]. As in the case of upper lip, intercepts given by these linear regressions cannot predict accurately enough the mean duration of the simple gesture (Jean-Luc: 129 *vs.* 158 ms; Annie: 90 *vs.* 161 ms; Benny: 85 *vs.* 148 ms; Christophe: 67 *vs.* 107 ms) and so the piecewise fitting reveals generally more appropriate.
If we consider now the slopes, it is also clear that only one speaker (Annie, the same as for upper lip protrusion) approaches the LA model (with 0.93), the three others behaving in rather close individual range (between 0.69 and 0.79), higher than for upper lip slopes (Fig. 1), but still not in the orthodoxy of LA (not to speak of TL).
If we want to give a schemata of these results, the only main difference with upper lip protrusion behaviour (Fig. 1), stays simply in the fact that the newly processed speaker (Christophe) has a rather small 100 ms duration for his basic contriction gesture [iy], compared with the 150-160 ms durations for the three others. But this is not a problem for our

model since the MEM specifies for each speaker his basic gesture values, then calculates every expansion knowing his expansion function, that can be fairly obtained with some test sample, manipulating *OI* from one consonant (about 100 ms for all speakers) to three or more (the maximum *OI* value depending on each speaker's rate habits: under 300 ms or up to about 400 ms).

4.1. Constriction phasing in *OI*

It is time now to set these results in relation to the acoustics, chosing a common reference event, to test if the procedure we used for upper lip protrusion is viable for lip constriction. To make short we will give only one example, Jean-Luc, knowing that the expansion functions we gave on Fig. 2 offer the possibility to calculate the fitness of the data of each speaker [1].

In Fig. 3a we represented, for this speaker, only the upper lip protrusion kinematic event *PO* (for *Protrusion Onset*), with the offset of [i] (*VVT[i]*) as the reference event (lower horizontal line at 0%; onset of [y], *VVO[y]* is the horizontal line at 100%). *%(PO-VVT[i])/OI* provides thus a *relative timing* measurement, say *phasing*.

In Fig. 3b, we did the same for area changes, using *10%Area.On* as a landmark comparable to *PO*, with the same reference event *VVT[i]*.

How does anticipation of these two events behave ? For movement onset (*PO*), it is clear that data point dispersion adopts a hyperbolic function (for this speaker as for others high regression coefficients were obtained with this fitting, from 0.82 to 0.93 [1]). The onset of protrusion can occur relatively well *into* the vowel [i] for small *OI* values (one consonant); and clearly *outside* of it (for *OI* values above 300 ms, corresponding here mainly to five consonants). We observe the same trend, with relatively less amplitude, for the constriction beginning event *10%Area.On*.

So we can say that the MEM holds for the two main components of rounding, protrusion and constriction.

5. DISCUSSION

Our Movement Expansion Model succeeds in accounting for the behaviour of the four French speakers under examination. MEM specifies for each a basic duration for the protrusion and constriction components of rounding, as well as an expansion function with a speaker-specific parameterization.

The fact that expansion coefficients vary between speakers may be reminiscent of a more abstract view of variation in language, i.e. the so-called "principles and parameters" approach in Chomsky's Universal Grammar. But in our concrete measurements this means simply that subjects follow globally and coherently the same expansion "law", with subject-specific parameters.

So to speak: vocalic gestures expand when they have temporal room enough between each other, regularly and at each speaker's own rate, without any "obligatory principle" urging them to fill between-vowel interval. This is a fairly different conception from both the look-ahead and time-locked ones.

Further work is in progress to test the MEM with the two other main components of vowel gestures: high-low and front-back dimensions.

Acknowledgements: This work was done in the frame of Esprit Basic Research project n°6975 *Speech Maps*.

REFERENCES

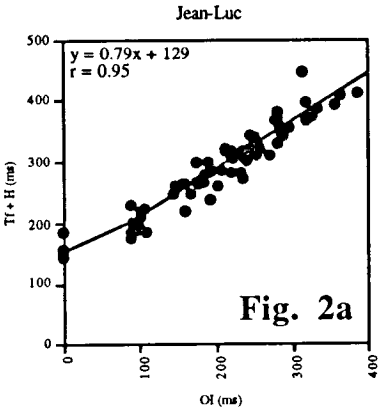
[1] Abry C. & Lallouache M.T. (in press). Pour un modèle d'anticipation dépendant du locuteur. Données sur l'arrondissement en français, *Bulletin de la Communication Parlée*.

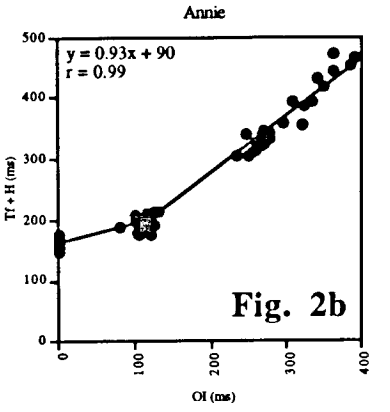
[2] Perkell J.S. & Matthies M.L. (1992). Temporal measures of anticipatory labial coarticulation for the vowel /u/ : Within- and cross-subject variability, *Journal of the Acoustical Society of America, 91*, 2911-2925.

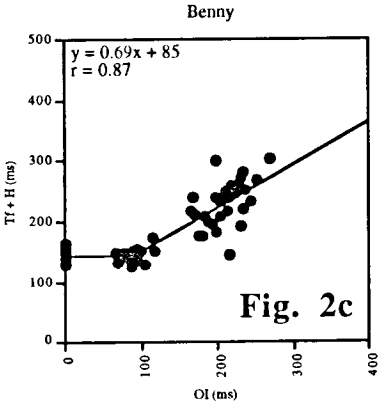
[3] Badin P., Motoki K., Miki N., Ritterhaus D. & Lallouache T.M. (1994). Some geometric and acoustic properties of the lip horn, *Journal of the Acoustical Society of Japan (E) 15, 4*, 243-253.

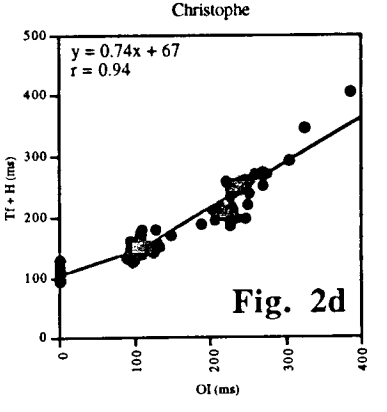
[4] Lallouache M.T. (1991). *Un poste «Visage-Parole» couleur. Acquisition et traitement automatique des contours des lèvres*, Thèse de l'INP, Grenoble.

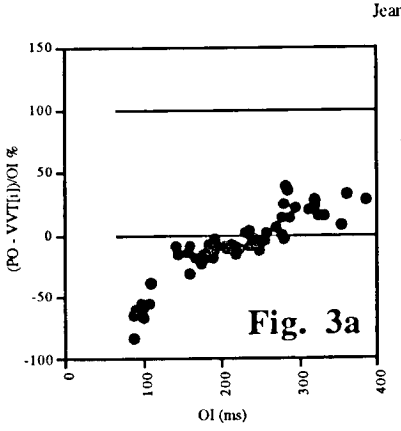
[5] Benoît C. (1986). Note on the use of correlations in speech timing, *Journal of the Acoustical Society of America, 80*, 1846-1849.

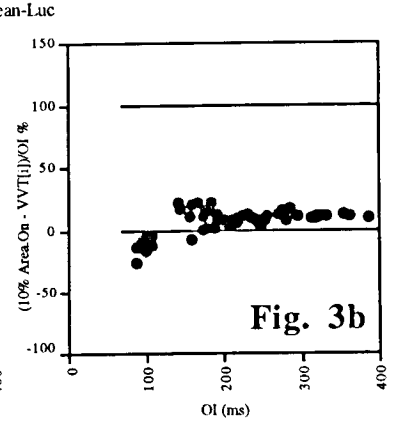












AUTOMATIC DERIVATION OF PHONETIC RULES BY AN ITERATED NORMALISATION PROCEDURE

Jesper Högberg
Department of Speech Communication and Music Acoustics, KTH, Stockholm, Sweden

ABSTRACT

In this paper an iterative normalisation procedure to automatically derive phonetic rules from a labelled speech corpus is described. It is assumed that the acoustic influence of coarticulatory constraints can be superimposed to model natural spectral variation. The algorithm proves to be promising when used to analyse the effect of phonetic context, stress and duration of Swedish front vowels on F1 and F2.

INTRODUCTION

Phonetic spectral variability in the realisation of a phoneme is due to numerous factors such as context, stress, speaker, speaking style, etc. Analyses in studies of coarticulation have traditionally dealt with syllables or words in a strictly controlled context. However the interaction between different factors in connected speech which influence segmental quality is very complex. In an attempt to describe essential coarticulatory phenomena we propose a data-driven method applied to a labelled speech corpus. The description is given in terms of a set of allophones or phonetic rules adjusting the spectral parameters of the phone to be modelled [1][2].

In this paper we describe a step by step normalisation procedure to automatically derive phonetic rules. The rules are easily interpreted and can be applied directly in the KTH text-to-speech system [3]. Thus, we combine the strengths of data driven and knowledge based techniques. The aim is to both produce more natural-sounding synthetic speech and also to gain deeper knowledge about speech production and perception.

In the current experiment we address the problem of modelling how F1 and F2 of Swedish front vowels are influenced by phonemic context, lexical stress and position. The variations along the speaker and style dimensions have been reduced by analysing read speech of one speaker.

METHOD

The speech material

The speech material consists of 11 short stories read by one male speaker. Formant frequencies of 2944 Swedish front vowels were manually measured. See Table 1 for the vowel distribution. This material has been used in several other investigations, e.g. [1][4]. Carlson & Nord [2] have also used the corpus to study context dependencies for the short vowel /e/.

Table 1. Number of analysed phonemes.

a	872	ø:	25
ʉ	83	ø	33
i:	184	e:	164
ɪ	355	e	647
y:	28	æ:	157
ʏ	49	æ	54
œ:	63	ɛ:	28
œ	37	ɛ	165

Derivation of rules

A data sample in the analysis consists of a prediction vector, ***X***, and a response vector ***Y***. The aim is to correctly predict ***Y***=[F1, F2], of a front vowel given ***X*** which contains information about the vowel's duration, lexical stress and phonemic context.

The phonemic context is defined by the identity of the target phoneme itself, the three preceding and the three following phones. Each sample also includes information about whether it is word final or word initial.

The algorithm is based on the assumption that the acoustic realisation of a phoneme can be modelled by superimposing the influence of the most important predictor variables.

A superpositional model has proven to be reasonably reliable despite statistically significant predictor variable interaction [5]. Context normalisation techniques have also been applied with success in automatic speech recognition, e.g. [6].

The samples are subjected to binary questions to find the group of samples that minimises the acoustic spread of the entire data set when those samples have been normalised and replaced. The amount of spread is evaluated by means of the function S,

$$S(y) = \sum_{i=1}^{2} \frac{1}{N} \sum_{j=1}^{N} \left(y_{i,j} - \bar{y}_i \right)^2$$

where $y_{i,j}$ and $\bar{y}_i$ are sample number j and the mean of the i:th formant frequency respectively. N is the total number of samples. All frequencies are calculated on the technical mel-scale. Hence, S is basically the sum of the formant frequency variances in mel.

A significant advantage of the replacement procedure is that all data are available for analysis in every iteration.

A categorical variable is a variable taking on unordered values, A question on such a variable can be of the type *"Is the phone immediately following the target a nasal?"* That is, phonetically meaningful features are used to form questions as well as single phoneme identities. Ideally, all phoneme combinations should be used to form questions. However, this task becomes unfeasible as the number of combinations, n, is given by $n=2^m$, where m is the number of phonemes. Currently 42 features are used apart from the single phoneme identities. A typical question on a continuous variable is *"Is the duration of the target phoneme < 100 ms?"* Questions are made on all unique target phoneme durations occurring in the speech corpus.

Samples responding positively to the question are normalised on the mel-scale towards the grand mean. The normalisation term that is added to the sample is the difference between the mean frequency of all samples and the mean frequency of the selected samples. One normalisation term is used for each formant frequency. The question which minimises the variance of the entire data set, in combination with the corresponding normalisation terms, is chosen to specify a rule.

A cross validation procedure is used to determine how many rules can be used without loss of predictive power for unseen data. Thus, for V-fold cross validation, (N/V)*(V-1) parts of the material is used for training and the remaining part is used for testing. The material is permuted so that each sample is used both for training and testing. The test score is calculated using the function S. The value of S, when applied in testing, is expected to decrease with increasing training until a critical point where the effect of overtraining will become noticeable and the variance will increase again. In the last step all data are used to generate rules. The cross validation result indicates the maximum appropriate number of rules that can be used without loss of generality.

In the experiments described below five-fold cross validation was used and no rule applying to less than ten samples was accepted. Moreover, all standard deviations are calculated on the mel-scale.

RESULTS

The overlap is considerable in the F1-F2 vowel space. In the first experiment, we employ the phonetic label of the target phoneme as a feature. Thus the predictive power of the phonetic labels can be compared to that of other features. All front vowels were analysed simultaneously and normalised towards a single front vowel prototype. F1 and F2 of this vowel, the grand mean, equal 514 and 1599 Hz respectively. The algorithm was iterated to generate 200 rules.

The normalised value of S as a function of the number of rules is plotted in Figure 1. The solid line indicates the mean cross validation score and the dashed line represents the result of the training on the entire data set.

The cross validation score indicates that no further improvement will be gained using more than about 50 rules. At this point 50% of the standard deviation for F1 and 52% for F2 is explained. As expected, the most significant rules concern the target

phoneme itself in terms of features. In fact, one third of the first fifty rules are of this type. The second most important factor influencing F1 and F2 is velar context. Rules number five and six concern front vowels in the immediate context of velar phones.

The distribution of rules based on the right and left context is quite symmetric. The exception is the far context, three phones away, in which the right context seems to be somewhat more important than the left context.

Only one rule among the first fifty, concerns duration or stress. This is quite natural since these aspects influence the target samples differently depending on their phonemic identity. Therefore, the question set was expanded to include composite questions such as *'Is the target /a/ AND stressed?'* Apart from simple phoneme identities, 16 additional features were used for the targets implying a dramatic increase in computational load.

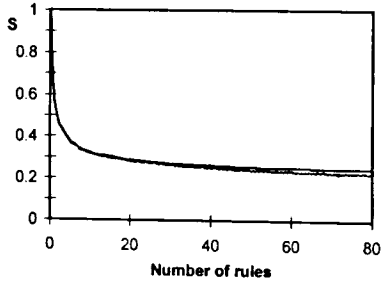


Figure 1. The value of the spread, S, as a function of the number of rules. The solid line indicates the mean cross validation score. The dotted (lower) line indicates the score from the training of the entire material.

The standard deviation of F2 is reduced from 162 to 72 mel using the first 50 composite rules. This explains more than 55% of the standard deviation. The standard deviation of F1 is practically unchanged. Thus, the introduction of composite questions yield only a slight improvement. However, the improvement seems to become more important when more rules are used. The cross validation score is one percent lower in the composite case when 50 rules are used.

A separate set of rules was generated for /a/, the most frequently occurring front vowel, to illustrate the power of the method more clearly and to extend the analysis to some extent without increasing the computational complexity.

The cross validation score indicates that some thirty rules suffice to model the major contextual influence on /a/. Degeneration occurs when more than fifty rules are used. The most significant factor is, again, velar context followed by lexical stress. The following two rules describe the coarticulatory influence of bilabials and nasals in the close vicinity of the target vowel. Vowel features are also important: rule number five and six consider /a/ coarticulated with other front and low vowels. The first 30 rules explain 27% and 39% of the standard deviation for F1 and F2 of /a/ respectively. This corresponds to a decrease from 48 to 35 mel in F1 and from 107 to 65 mel in F2. More rules are based on the left context than on the right among the top thirty rules. This means that the left context has stronger predictive power than the right. Moreover, the stronger explanatory power of the left context mainly concerns F2. It is unclear whether this has any implications for reasoning about carry-over vs. anticipatory coarticulation before the phonetic distribution of the context is analysed more thoroughly.

Since only one speaker, reading text passages, is analysed we expect the articulatory effort and speaking style to be about the same throughout the speech material. It is reasonable to believe that the duration of the target phoneme will have a systematic effect on the formant frequency values [7]. Therefore, when the best rule has been found a duration-dependent normalisation adjustment is introduced to refine the analysis. The formant frequency displacement is assumed to increase linearly with a logarithmic decrease in segment duration.

Figure 2 shows an example of the relation between the second formant frequency of /a/ tokens following immediately after a velar segment and the logarithmic value of the duration.

There was a decrease in the standard deviation of both F1 and F2 on the training of the entire data set. The mean cross validation score indicates a small improvement compared to the normalisation independent of segment duration.

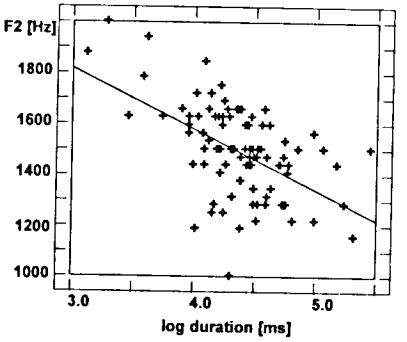


Figure 2. F2 of /a/ following a velar segment plotted vs. the logarithmic value of the duration. r = -0.54.

DISCUSSION

In this paper we have proposed a method for automatic derivation of phonetic rules from a labelled speech corpus. In the first experiment the phonetic labels assigned to the target vowel proved to be powerful formant frequency predictors as expected. However, all front vowel identities were not more important than the context. This might change if F3 is taken into account as well.

The cross validation score plotted in Figure 1 implies that the method is robust. The curve does not turn upwards, indicating over-fitting to test data, not even after 200 iterations. The cross validation of the /a/ rules displays a slight deterioration if more than 60 rules are used. The robustness is probably due to the fact that the overall decrease in variance depends on more than the magnitude of a coarticulatory effect. Just as important is the number of samples influenced. The rules were used to predict formant frequencies given the predictor vectors of the training material for /a/. The result showed that there is a systematic tendency of underestimating high formant frequencies and overestimating low frequencies. That is, the formant frequency displacement seems to be underestimated on an average. One plausible explanation for this is that the normalisation terms are based on differences in mean values that are biased by other coarticulatory effects. We conclude that the algorithm proposed in this paper is robust and provides easily interpretable results that potentially can be used to enhance the quality of synthetic speech.

ACKNOWLEDGEMENTS

I would like to thank Rolf Carlson for initiating this work and for great moral and practical support. This work has been supported by grants from The Swedish National Language Technology Programme.

REFERENCES

[1] Högberg, J (1994), A phonetic investigation using binary regression trees. In: Papers from the Eighth Swedish Phonetics Conference, Lund, Sweden.

[2] Carlson R & Nord L (1993), Vowel dynamics in a text to speech system - some considerations. In: Proc. of Eurospeech 93, 1911-1914.

[3] Carlson R, Granström B & Hunnicutt S (1991). Multilingual text-to-speech development and applications. In: (Ainsworth AW, ed.), *Advances in speech, hearing and language processing*, London: JAI Press, UK.

[4] Neovius L & Raghavendra P (1993), Comprehension of KTH text-to-speech with 'listening speed' paradigm. In: Proc. of Eurospeech 93, 1687-1690.

[5] Broad DJ & Fertig RH (1970) Formant-frequency trajectories in selected CVC-syllable nuclei. In: J.Acoust SocAm, Vol 46, 1572-1582.

[6] Philips M, Glass J & Zue V (1991) Automatic learning of lexical representaions for sub-word unit based speech recognition systems. In: Proc European Conf. on Speech Comm. and Technology.

[7] Lindblom B (1963) Spectrographic study of vowel reduction. In: J.Acoust SocAm Vol 35, 1773-1781.

A CROSS-LINGUISTIC STUDY OF BLENDING PROCESSES IN CONNECTED SPEECH

M.J. Solé and E. Estebas
Laboratori de Fonètica, Universitat Autònoma de Barcelona, Spain

ABSTRACT

Assimilatory and blending processes in connected speech in English and Catalan in a variety of conditions (slow vs fast speech, functional vs lexical words and oral vs nasal segments) were studied. The intergestural adaptation of alveolar+dental/palatal clusters were analyzed with simultaneous EPG, acoustic and EGG data. The results show gradient blending processes rather than discrete substitution of features. Sliding trajectories of two strongly coarticulated gestures and adaptation of the constriction location of C1 to that of C2 were found. The interactions betwen the two consonants in the cluster were significantly affected by the oral/nasal nature of C1 and to a lesser extent by rate and word category. The observed crosslinguistic differences require an explanation of blending in terms of language dependent structural factors.

INTRODUCTION

This study investigates assimilatory and blending processes in connected speech in English and Catalan in the light of current phonological theories. The modelling of connected speech processes has particular relevance for speech technology and for developing models of speech production.

Blending has traditionally been represented in phonological analyses as a categorical process (expressed in terms of phonological rules or feature linking processes) by which some feature of the segment is replaced by another feature. Such view assumes a discrete change in the input to speech production and in the neuromuscular commands to articulate the segment. Models of speech production and gestural phonology, on the other hand, claim that blending processes involve gradient modifications resulting from overlap of two competing contiguous gestures involving the same articulator, with no change in the input to speech production. Such overlap will show intergestural accomodation giving rise to a shift in the articulatory configuration of C1 only, a sliding movement of two strongly coarticulated gestures or one single gesture for the cluster with an intermediate target. Articulatory and acoustic data on blending in connected speech [1, 2] show residual alveolar gestures for the alveolar provide support for the gradient nature of assimilatory processes. In this paper the articulatory behaviour of blending processes -- i.e., those involving two successive consonantal gestures produced with the same articulator -- was examined in English and Catalan across different speech rates. The aim of the study was to determine 1) whether blending processes involve a planned reorganization of articulatory movements or the modification, due to overlapped instructions and simultaneous articulation, of the individual trajectories of C1 and C2, and 2) the gradient vs categorical behaviour of blending processes in a variety of conditions: slow vs fast speech, oral vs nasal alveolars, functional (e.g. *on thoughts*) vs lexical (e.g. *John thought*) words and English vs Catalan.

METHOD

Simultaneous EPG, acoustic and electroglottographic data were obtained for six repetitions of oral/nasal alveolar consonants that can occur in word-final position in these languages followed by a word-initial consonant involving the same articulator (VC1#C2V). Comparable VC1#V and V#C2V utterances were also analyzed. Low vowel contexts were chosen to avoid anterior contacts due to high vowels. The interactions between C1 and C2 were studied in the following clusters: /n/+/tʃ/ in English and Catalan, /t/+/tʃ/ in Catalan and /d/+/tʃ/ in English (in Catalan only voiceless obstruents are possible in word final position; in English word final /d/ was chosen to avoid (pre)glottalization of final /t/), /n/ + /t/ in Catalan (apicoalveolar + apicodental stop) and /n, d/+/θ/ in English. The test sequences involved meaningful phrases spoken by two native speakers each of English (speakers AR and MM) and Catalan (speakers MJ and AF) at slow and fast rate. Overall 1056 tokens were analyzed. Variation in the degree and distribution of linguopalatal contact in the two consonants in the cluster was analyzed using the EPG contact index method [3] and compared to that of canonical VC1#V and V#C2V utterances. For each consonant the constriction location and constriction degree was assessed by means of the contact anteriority (CA) index (since all consonants involved are produced in the dento-palatal zone), and the contact centrality (CC) index. CA reflects variation in the degree of contact fronting and CC reflects variation in the degree of contact from both sides of the palate towards its median line. The value of the two indices increases as the linguopalatal contact area becomes larger or more anterior (CA), or approaches the central zone or shows a larger central contact area (CC). The contact indices were calculated for the EPG frames showing maximum contact (PMC).

To examine whether the modification of the articulatory configuration of consonants in the cluster is the result of blending of the two gestures, or rather an assimilatory process we compared the CA and CC indeces at the PMC in C1 (and C2) in the cluster with those of intervocalic C1 (and C2), and the CA and CC indeces of C1 in the cluster with those of C2 in the cluster. Single/cluster X Rate (and Word type) ANOVAS were performed for each speaker and each cluster type separately in order to find out whether single consonants and consonants in clusters showed significant differences in contact indeces, and whether significant effects were rate and word category dependent. C1/C2 X Rate (and Word category) ANOVAS were performed in order to determine whether the differences in contact indeces between C1 and C2 in the cluster were significant, and dependent on rate or word category.

RESULTS

First we will comment on differences in linguopalatal contact indeces for C1 (and C2) in isolation and in clusters, and then we will analyze whether such differences lead to blending of the the two individual gestures in clusters. Finally rate and word category effects will be analyzed for each cluster type.

English /nθ/ and /dθ/ clusters

For the two speakers, /n, d/ in the cluster show a significant increase in CA values vis-à-vis intervocalic position, which indicates that the alveolar segment becomes more anterior due to the effect of dental /θ/. Speaker AR also shows a significant decrease in central contacts (CC) for /n/ in the cluster (only in function words) making the constriction degree of this segment more similar to that of fricative /θ/, which shows virtually no central contacts intervocalically. Speaker MM, on the other hand, shows no variation in CC indeces for the first segment in the cluster, but the central contacts for the second segment are significantly higher than in singleton /θ/, indicating that the constriction degree of C2 becomes more similar to that of the alveolar stop. Thus, the adaptation of the constriction degree for both segments in the cluster go in opposite direcions for the two speakers. CA values of /θ/ preceded by /n, d/ increase significantly vis-à-vis intervocalic position indicating a more constricted articulation of /θ/ in clusters. Although the articulatory configurations of both segments in the cluster are modified, becoming more similar to each other, there seems to be no blending of the constriction location or constriction degree of original /n, d/ and /θ/ since the CA and CC values of /θ/ in the cluster are significantly lower than those for the first segment. Alternatively, the lower CA values of /θ/ may reflect the lesser degree of contact for fricatives with blending of the constriction location for the cluster at the dental region.

Rate effects are speaker dependent. Speaker MM shows no significant effect of rate. For speaker AR differences in rate affect the contact indeces of C1 and C2 in the cluster showing more adaptation of C1 to C2 in fast than in slow speech. Word category effects show that C1 (and C2) in lexical words tend to have a higher constriction degree than in function words both in isolation and in clusters.

Catalan /nt/ clusters

Compared to intervocalic /n/, /n/ followed by dental /t/ shows a significantly higher CA and CC values for both subjects. The second consonant in the cluster shows a smaller and less consistent increase in both indeces. Such increase in anteriority and centrality in both segments in the cluster indicates an extension in the area of closure, that is, a larger spatial displacement of the tongue tip/blade for the two similar lingual gestures. Thus, coarticulation between the two same-tier gestures has the effect of reinforcing each individual gesture (similar reinforcement effects in clusters, resulting in a generally more advanced and higher maximum position, were found by [4, 5]). As a result of reinforcement, which affects C1 in a

larger degree than C2, the significant difference in CA indeces found at the two points of maximum constriction for /n/ and /t/ intervocalically disappears at the corresponding points in the cluster, which is articulated as a single gesture in the region of dental /t/.Thus, /nt/ trajectories in Catalan show shifting of the constriction location of /n/ to that of dental /t/.

No significant effects of rate and word category were found except for speaker MJ who exhibits significantly more similar central contacts for the two consonants in the cluster in functional words in fast speech.

Catalan /n(t)ʃ/ /t(t)ʃ/ clusters and English /ntʃ/, /dtʃ/ clusters

In Catalan /ʃ/and /tʃ/ are in free variation in word initial position. Speaker AF consistently produced [tʃ] intervocalically and in clusters whereas speaker MJ consistently produced [ʃ]. Thus, the CC index for [ʃ] for speaker MJ is much lower, indicating an open constriction, than that for AF's [tʃ]. Compared to intervocalic /n, t/, Catalan /n, t/ followed by an alveolopalatal obstruent show a significantly more retracted constriction location (lower CA), indicating that the articulatory configuration of C1 becomes more similar to that of C2. No significant differences in constriction location between cluster and intervocalic C2 were found, indicating that all the coarticulatory effects go in the anticipatory direction (these results are in line with those found for Italian [4]). The CC values of C1 and C2 in oral and /n/ clusters increase vis-à-vis intervocalic position, indicating that the constriction degree is higher for both segments in the clusters than in single consonants, thus confirming the effect of reinforcement suggested above.

For speaker MJ, the difference in CA between the two segments in isolation disappears in the oral and /n/ clusters. This suggests blending of the individual gestures constriction location. The cluster appears as a single gesture realized in a region in the vicinity of single /ʃ/. Whereas /n/ clusters show blending of constriction degree (no difference in CC between the two elements in the cluster) with /ʃ/ adapting to the constriction degree of single /n/, /t/ clusters exhibit a significant discontinuity in the central contacts for the two elements, indicating the transition from a stop to a fricative degree of constriction. For speaker AF there is no blending of the individual gestures constriction location or degree (significantly different CA and CC in the two segments in the cluster).

For the English speakers both consonants in the cluster show a higher CA index than in intervocalic position, indicating that the extent of contact is higher in the cluster. In all cases, the constriction location of C2 in the cluster becomes more similar to that of single C1 (carry over assimilation). The alveolar segment in the cluster shows a much wider contact in the central region (significantly higher CC) due to the effect of the following alveolopalatal (anticipatory assimilation). No changes in central contacts for the /tʃ/ are observed in the cluster. Thus, there is a mutual influence of both segments in the cluster: the first segment adapts to the wider central constriction of the second element whereas the /tʃ/ adapts to the constriction location of the alveolar.

For speaker MM the CA values of /d/+/tʃ/ differ significantly whereas those of /n/+/tʃ/ do not differ (showing values in the region of single /n/). The CC values for both elements in the cluster, however, differ significantly, suggesting two different gestures. This confirms the visual impression that the cluster is produced with two distinct but strongly coarticulated /n,d/ + /tʃ/ gestures. Speaker AR shows no blending of the individual gestures.

Overall, fast rate tends to exhibit lower CA and CC indices than slow rate both in clusters and in isolation, indicating a smaller spatial elevation of the articulator in fast speech. However, rate differences seem to be speaker dependent (English speaker MM shows no significant effect of rate on contact indices). All consonants (C1 and C2), and specifically oral alveolars, tend to show a wider extent of contact in lexical than in functional words both in clusters and in isolation. Comparison of the contact index values for both consonants in the cluster at slow and fast rate and in lexical and functional words shows no significant interaction effects. Thus, rate and word category do not affect the degree of blending in clusters.

It is necessary to be cautious when drawing conclusions about differences in blending processes between languages in view of the small number of speakers and the differences between speakers within each language group. However, it can be observed that in alveolar + alveolopalatal clusters English shows a mutual influence of both segments in the cluster and no cases

of articulatory shift whereas Catalan speakers tend to adapt the articulatory configuration of C1 to that of C2, the latter dominating the articulatory configuration of the cluster. This observation is in line with the stronger tendency in Catalan to weaken coda consonants, and to show a greater tendency to articulatory overlap than English and other languages [6].

CONCLUSIONS

The results indicate that featural phonology cannot fully account for the data obtained. Evidence of the alveolar gesture suggests that the alveolar segment is present in the input to speech production although its realization may be modified in connected speech. Thus, assimilatory processes cannot be modelled in terms of substitution of features. Articulatory Phonology predicts than when two same-tier gestures overlap in time, blending of the constriction locations of C1 and C2 result in one single gesture with an intermediate target [1]. Furthermore, blending of constriction location is claimed to occur only when the specifications for constriction degree in both segments in the cluster are the same [5]. In this view blending is interpreted as a reorganization of the articulatory movements at the gestural level, and not at the tract variable level. These predictions are not fully borne-out by the data obtained, which show no cases of intermediate targets for the blended gestures. In cases of articulatory shift the constriction location of C1 in the cluster is modified in the direction of that of C2. When two separate trajectories are found, there is mutual adaptation of the constriction location of C1 and C2. As regards the relationship between constriction location and constriction degree in blending, the predictions of AP are only partially borne out. Blending of constriction location occurs in Catalan /n/+/t/ clusters, which share constriction degree, but also possibly in English /n,d/+/θ/ clusters involving different specifications for constriction degree. In alveolar + alveolopalatal clusters, blending occurs for Catalan /n, t/ +/ʃ/, involving two different constriction degrees whereas it does not occur in clusters involving /tʃ/.

The rate, word category and speaker dependency in blending processes, along with the predominance of C2 constriction location, suggest that the neural commands for the alveolar segment are overlapped and modified by the conflicting commands for the upcoming segments, rather than reorganized and modified at a higher gestural level. The language-specific nature of blending processes and the higher occurrence of articulatory shift in Catalan than in English does not allow an explanation in terms of the organization of articulatory gestures in fast speech alone -- as claimed by gestural phonology -- but requires an explanation in terms of language dependent structural factors. The language-specific ranking of universal constraints, as proposed by Optimality Theory, can explain the observed cross-language differences.

The fact that nasal alveolars are more likely targets of place assimilation than oral alveolars can be explained in terms of trade-offs between articulatory effort and enhanced perceptibility: oral stops have stronger place cues than nasals and are more likely preserved in articulation than acoustically less salient segments.

ACKNOWLEDGEMENTS

This research was supported by a DGICYT grant PB93-859 to the Universitat Autònoma de Barcelona.

REFERENCES

[1] Browman, C.P. and L.M. Goldstein (1990), Tiers in Articulatory Phonology. In J. Kingston and M. E. Beckman (eds.), *Papers in Laboratory Phonology I.* Cambridge University Press: Cambridge, pp. 341-376.

[2] Nolan, F. 1992. The descriptive role of segments: Evidence from assimilation. In G.J. Docherty and D.R. Ladd (eds.), *Papers in Laboratory Phonology II.* C.U.P.: Cambridge, 261-280.

[3] Fontdevila, J., M.D. Pallarès and D. Recasens. 1994. The contact index method of EPG data reduction. *Journal of Phonetics*, 22, 141-154.

[4] Farnetani, E. and M.G. Busà. 1994. Italian clusters in continuous speech. *Proceedings of the 3rd International Conference on Spoken Language Processing*, Vol. 1, pp. 359-362.

[5] Romero, J. in press. Articulatory blending of lingual gestures. *Proceedings of the 1993 ACCOR Meeting*. Barcelona.

[6] Gibbon, F., W. Hardcastle and K. Nikolaidis. 1993. Temporal and spatial aspects of lingual coarticulation in /kl/ sequences: a cross-linguistic investigation. *Language and Speech*, 36 (2, 3), 261-277.

EFFECT OF VOICE QUALITY ON THE TENSE/LAX DISTINCTION FOR ENGLISH VOWELS

Keith R. Kluender, Andrew J. Lotto, and Lori L. Holt
Dept. of Psychology, University of Wisconsin-Madison, USA

ABSTRACT

For a variety of East and West African languages, voice quality covaries with tongue root advancement. This regularity may be due to the mutually enhancing auditory effects of breathy voice and a low first formant frequency. Evidence is adduced for this explanation in results from perceptual categorization experiments in which voice quality and formant values were independently manipulated.

I. INTRODUCTION

In a number of vowel systems, there exists an articulatory covariation between advanced tongue root as a distinctive feature and voice quality. For a variety of East and West African languages, an advanced tongue root vowel is produced with a breathy phonation (sometimes called 'lax' voice); whereas, a modal or even creaky voice quality is used with non-advanced tongue root [1,2,3]. In fact, for some vowel harmony systems such as Akan, breathy voice has been referred to as, "the main auditory correlate of root advancing" [3].

Covariations between presumably distinctive articulatory variables often suggests that the variables are mutually enhancing either in terms of articulatory ease or perceptual distinctiveness. The experiments reported here evaluate a potential explanation for this regularity based on hypothesized perceptual advantages from the interaction between acoustic effects of vocal-tract shape and voice quality.

The acoustic effects of breathy phonation on the resulting vowel are essentially twofold. Due to the longer open quotient of the glottis, which characterizes breathy phonation, the resultant waveform has relatively greater energy at low frequencies. The amplitude for the fundamental component (H1) increases and the falloff across higher frequencies is steeper [4]. Similarly, the relatively low first formant of [+advanced tongue root] high vowels contributes to greater energy at low frequencies. The hypothesis being considered is that the joint acoustic consequence of vocal-tract shape and voice quality interact in a perceptually enhancing manner by jointly contributing to a low frequency prominence.

Because it has been suggested that the advanced tongue root contrast is similar to the tense/lax distinction in English [5], perception of English vowels may provide a reasonable experimental test of this hypothesis. The contrast between tense and lax high vowels in English (e.g. /i/ vs. /ɪ/) is signalled, in part, by the lower frequency of the first formant (F1) in tense vowels. If these vowels are produced with a breathy voice, it is possible that the increased H1 and steeper dropoff will lead to a lower perceived F1 and hence to a more 'tense' vowel. The following experiments tested this possibility by obtaining categorizations from native English speakers for tense and lax high vowels varying in voice quality.

II. EXPERIMENT 1

Subjects.

Fifteen college-age adults, all of whom learned English as their first language, served as listeners. All reported normal hearing. Subjects received Introductory Psychology course

results of Experiment 1 is that the increased spectral tilt of the series modeled after breathy productions degraded the higher frequencies to the point where subjects were forced to rely solely on the first formant for identification. This is especially a possibility for the back vowel series, since a reasonable sounding /u/ can be constructed from a single low formant. Experiment 2 was designed to parcel out the effects of exaggerated spectral tilt and the increasing of H1 amplitude.

Subjects.

Subjects were eleven college-age students, none of whom had participated in Experiment 1. All subjects reported normal hearing and learned English as their first language. Course credit was awarded for participation.

Stimuli.

The stimuli were identical to those used in Experiment 1 except that for the "breathy" series only the amplitude of H1 (OQ = 72) was increased. Spectral tilt remained the same for all series.

Procedure.

The procedure was identical to that used in Experiment 1.

Results.

Again, identification boundaries were calculated using probit analysis. The mean boundary values are displayed in Figure 2. As in Experiment 1, breathy phonation led to more vowels being identified as tense for series modelled after female productions. However, in the case of vowels modelled after male productions, breathy phonation, synthesized solely through manipulation of H1 amplitude, did not lead to more vowels being identified as tense.

These results suggest that the amplitude of the first harmonic is largely responsible for the identification boundary shifts in vowels modelled after female productions, whereas overall spectral tilt seems to determine the effect produced by male series.

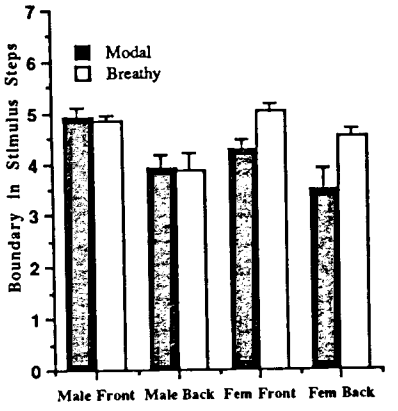


Figure 2. Probit identification boundary values (in stimulus steps) for the eight series in Experiment 2.

IV. DISCUSSION

The purpose of this investigation was to determine the effect of a potential perceptual interaction between voice quality and first formant frequency on the categorization of [+high] English vowels. The data indicate that the quality of voice used to produce a high vowel will affect whether the vowel is perceived as tense or lax.

These results suggest that the covariation in African languages may be consistent with general principles of Auditory Enhancement [6] and Adaptive Dispersion [7]. These theoretical frameworks predict that articulatory factors which tend to enhance the perceptual discriminability of a resultant speech sound will be favored in phonetic inventories. Breathy phonation enhances the low frequency prominence of tense high vowels and of [+advanced tongue root] high vowels. According to the present results, if a language included breathy production of lax vowels then the distinctiveness of the tense/lax contrast would be reduced. However, it

credit for their participation.

Stimuli.

Four vowel series were synthesized with eight endpoints modeled after male and female productions of /i/ and /ɪ/ (front), and /u/ and /ʊ/ (back). Five intermediate vowels between tense (/i/, /u/) and lax (/ɪ/, /ʊ/) were synthesized by manipulating the nominal center frequencies of the first three formants. In particular, the frequency of F1 varied linearly from 210 Hz to 330 Hz for the male front series and from 330 Hz to 440 Hz for the male back series. For the series based on female productions, F1 varied in equal steps from 310 Hz to 430 Hz for the front vowels and from 370 to 480 Hz for the back vowels. Fundamental frequency was 135 Hz for series modeled after the male productions and 233 Hz for the female series and all stimuli were 120 msec in duration. Breathy versions of each seven-step series were created by increasing the amplitude of the first harmonic (open quotient; OQ = 72) and increasing spectral tilt (TL = 17) using the Klatt software synthesizer [4].

Stimuli were synthesized with 12-bit resolution at a 10-kHz sampling rate and stored on computer disk. Stimulus presentation was under control of a microcomputer. Following D/A conversion (Ariel DSP-16), stimuli were low-pass filtered (Frequency Devices 677, cutoff frequency 4.8 kHz) prior to being amplified (Stewart HDA4), and played over headphones (Beyer DT-100) at 75 dB SPL.

Procedure.

Subjects participated in a two-response forced choice identification task arranged in randomized blocks with ten presentations per stimulus. Each block contained the breathy and modal versions of each gender x place of articulation series, for a total of four blocks (male/female x front/back). Stimuli were presented at a rate of approximately one stimulus every 3 seconds. Subjects identified each vowel by pressing the appropriate button on a response box connected to a microcomputer. The buttons were labeled 'beat' and 'bit' or 'boot' and 'book'.

Results.

For each of the eight series, the identification boundary value was determined for each subject using probit analysis. The mean boundaries for each of the eight series are displayed in Figure 1 in terms of the stimulus steps (one = 'lax' endpoint; seven = 'tense' endpoint).

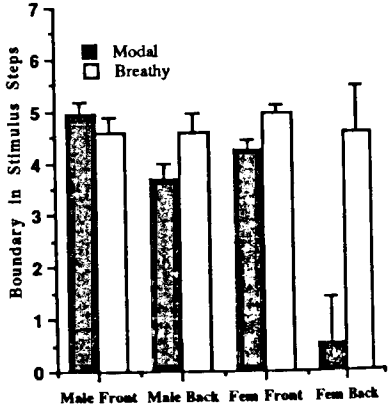


Figure 1. Probit identification boundary values (in stimulus steps) for the eight series of Experiment 1.

The results indicate that, in general, breathy phonation led to more vowels being identified as tense, particularly for male and female /u/-/ʊ/ series. Except for the male /i/-/ɪ/ series, the boundaries for the breathy series were significantly greater (more 'tense') than the boundaries for the modal series.

These results suggest that there exists a perceptual interaction between tenseness/laxness and phonation type.

III. EXPERIMENT 2

One possible explanation for the

appears that languages tend to contain the more distinctive pairing of breathy phonation with advanced tongue root and creaky voice (which would de-emphasize the low frequencies) with unadvanced tongue root.

There is also an interesting divergence in the data based on whether the stimuli were modeled on male or female productions. The effect of breathiness on tense/lax categorizations was much more robust for the female series. Even when spectral tilt was held constant in Experiment 2, subjects labeled the female vowels differentially when the amplitude of the first harmonic was changed.

It has been widely reported that females use breathy phonation more often than men in a variety of languages, including English [4,8]. If breathy productions lead to a more 'tense' high vowel, then it would be advantageous for English speaking females to use breathiness distinctively, i.e. use a breathy voice for tense high vowels and a modal voice for lax high vowels. Since the effect of voice quality is less robust in males, it is less probable that they would develop such a "strategy" for productions.

There is a dearth of relevant data concerning females' voice quality across the vowel space. It *has* been reported that females' produce a larger vowel space than males [9]. Even when differences in vocal tract size are accounted for, the variation in F1 across the space is larger for females. If females *are* producing high tense vowels with a breathy phonation, then the increased amplitude of the first harmonic could be effectively lowering F1 causing a non-linear stretch of the F1 space.

Along with the regularity among African languages, this could be another example of the premium placed on the perceptual distinctiveness of speech sounds.

This work was supported by NIDCD Grant DC-00719 and NSF Young Investigator Award DBS-9258482 to the first author.

IV. REFERENCES

[1] Berry, J. (1955), "Some notes on the phonology of the Nzema and Ahanta dialects", *Bulletin of the School of Oriental and African Studies*, vol. 17, pp. 160-165.

[2] Jacobson, L.C. (1980), "Vowel-quality harmony in Western Nilotic languages", in R.M. Vago (ed.) *Issues in Vowel Harmony*, John Benjamins: Amsterdam, pp. 183-200.

[3] Stewart, J.M. (1967), "Tongue root position in Akan vowel harmony", *Phonetica*, vol. 16, pp. 185-204.

[4] Klatt, D.H., & Klatt, L.C. (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *Journal of The Acoustical Society of America*, vol. 87, pp. 820-857.

[5] Halle, M., & Stevens, K. (1969), "On the feature 'Advanced Tongue Root'", *MIT Research Laboratory of Electronics Quarterly Progress Report*, vol. 94, pp. 209-215.

[6] Diehl, R.L., & Kluender, K.R. (1989), "On the objects of speech perception", *Ecological Psychology*, vol. 1, pp. 121-144.

[7] Liljencrants, J., & Lindblom, B. (1972), "Numerical simulation of vowel quality systems: The role of perceptual contrast", *Language*, vol. 48, pp. 839-862.

[8] Henton, C.G., & Bladon, R.A.W. (1985), "Breathiness in normal female speech: Inefficiency versus desirability", *Language & Communication*, vol. 5, pp. 221-227.

[9] Fant, G. (1975), "Non-uniform vowel normalization", *Quarterly Progress Status Report*, RIT Stockholm, vol 2/3, pp. 1-19.

FINNO-UGRIC PROSODIC SUBSTRATA IN THE GERMANIC LANGUAGES AND VICE VERSA

Kalevi Wiik
University of Turku, Finland

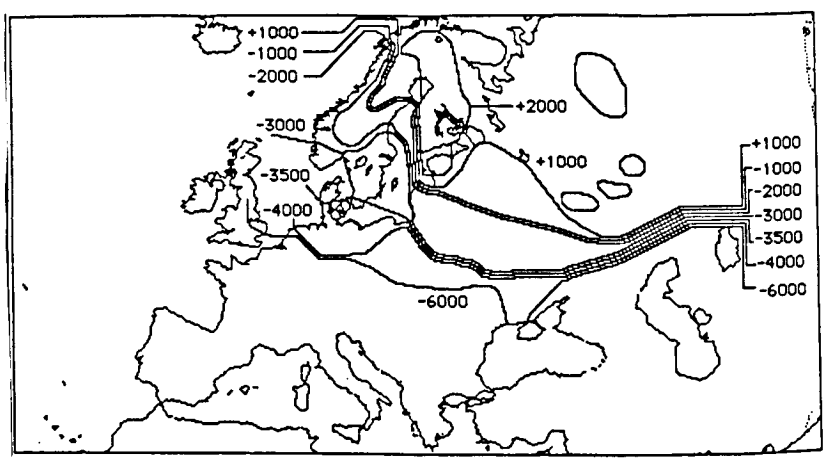
INTRODUCTION

The prosodic area around the Baltic Sea has long been of interest to linguists and phoneticians. The best known presentations of this *Sprachbund* are those of Roman Jakobson and Ilse Lehiste. In this article I return to the old questions and try to give a phonetically and archaeologically oriented solution to the question of how the people of quite different linguistic backgrounds make use of similar tonal and durational features in their speech.

ARCHAELOGICAL/HISTORICAL BACKGROUND

One of my starting points is the contention of some modern archaeologists (e.g. Paul Dolukhanov and Milton Nunez) that the languages of the populations of the periglacial zone of the latest Glacial Period in Europe were Uralic. The periglacial zone is roughly equal to the Baltic Sea basin. Accordingly, the original language of the Baltic Sea area was Uralic, later Finno-Ugric (FU) and later still Finnic. The Indo-European (IE) language arrived in this area with the spread of agriculture (Colin Renfrew's Model) and/or with the warlike Kurgans (Marija Gimbutas' Model). The approximate locations of the language boundary during the last eight thousand years are seen in the map below.

In addition to the gradual spreading of the IE area towards the north, there were a number of IE migrations to the northern and eastern coasts of the Baltic Sea: (1) the Baltic and partly Germanic population of the Battle Axe Culture ca 2500-2000 BC, (2) the Germanic population of the Scandinavian Bronze Culture ca 1400-900 BC, (3) the Scandinavian Iron Age population ca 200-400 AD, (4) the Scandinavian Vikings ca 800-1000 AD, (5) the Swedish population of the Crusade Period to the islands and coasts of Finland and Estonia at approximately 1000-1300 AD, and (6) the Low German and Baltic Hansa traders to the trading centers of Estonia and Latvia at approximately 1200-1600 AD.



TYPES OF LANGUAGE CONTACTS

The contacts have been of two types: either the FU people have gradually changed their language into an IE language, or the IE people have gradually changed theirs into a FU one. The former type of language shift (FU > IE) would have taken place at and near the FU-IE language boundary that has moved from northern Germany and Poland to northern Scandinavia and the southern border of Estonia. The latter type (IE > FU) would have taken place on the coastal areas of Finland, Estonia, and Latvia, where the IE settlers would have assimilated with the local indigenous FU people. The FU > IE type of language shift would have left a FU substratum in the IE languages, and the IE > FU shift would have left an IE substratum in the FU languages. As a result, the original, more or less homogeneous IE protolanguage of northern Europe was split into dialects (which later developed into independent languages) on the basis of the FU substrata left by the indigenous FU population. I believe this is the way Proto-Germanic (perhaps also Proto-Balto-Slavic) separated from Proto-Indo-European and how many of the dialect boundaries of the northern IE languages came about. On the norhtern and eastern coasts of the Baltic Sea, the Finnic languages were split into dialects according to the strength of the Germanic influence in them.

THE RELEVANT PROSODIC DIFFERENCES BETWEEN FU AND IE PROTOLANGUAGES

There are four relevant word prosodic differences between Proto-Indo-European and Proto-Finno-Ugric:
(1) free stress in IE, initial stress in FU,
(2) grave tone in IE, no grave tone in FU,
(3) foot isochrony in FU, no foot isochrony in IE, and
(4) vowel harmony in FU, no vowel harmony in IE.

By the term "grave tone" I mean a tonal peak that occurs on an unstressed syllable. The grave tone in the early phase of the development of the Germanic languages occurred on a "long" unstressed syllable. A "long" syllable is here a syllable that has a long vowel or a diphthong, or which ends in a consonant. "Foot isochrony" refers to the tendency of having each foot (a sequence of two syllables, sometimes of one syllable or three syllables) of approximately equal length. In particular, if the first syllable is long, the second syllable vowel (sometimes the whole second syllable) is short, and vice versa.

STRESS

The Finno-Ugrians learning the IE protolanguage and later some of its daughter languages met a difficulty in pronouncing primary stress on a non-initial syllable and particularly in knowing on which syllable stresss should be put in each particular word form. They spoke the IE language with the initial stress of their own FU language. The result was that (1) Proto-Indo-European was pronounced with initial stress in the area of northern Germany, Denmark, and Scania. In traditional IE linguistics this is the main event in the split of Proto-Indo-European into Proto-Germanic and the other IE languages. The Germanic *Akzentverschiebung* or the shift of stress to the root initial syllable is, no doubt, a Finno-Ugric substratum in the Germanic languages.

In connection with the clash of the two language systems and the resulting stress shift, the significance of the first (stressed) syllable increased at the cost of the non-first (unstressed) syllables, and also physically, more energy was now concentrated on the first syllable. This is an expected result: The physical manifestations of stress are usually more conspicuous in a language with phonological stress than in a language with only automatic (demarcative) stress. When stress was moved to the first syllable in Proto-Germanic, the physical manifestations of stress remained more or less unchanged; the change concerned only the place of stress. The result was what is sometimes called "stress enhancement" or "stress concentration". This in turn caused a large number of changes in the segmenal phonology of the language. One of the consequences was Verner's Law. The unstressed syllables with a grave accent were perceived to be "too strong", if there was a voiceless obstruent at the beginning of the syllable;

i.e. these syllables did not sound sufficiently unstressed in relation to the initial syllable. To avoid the sensation of a stress peak in a non-initial syllable, the initial voiceless obstruent of these syllables had to be weakened. This took place by weakening the syllable initial voiceless obstruent into a voiced one; in other words the phenomenon of Verner's Law. Later this phenemonen was also introduced into the Finnic languages, when Germanic settlers came to the coasts of Finland and Estonia. The phenomenon there is called consonant gradation (*astevaihtelu, astmevaheldus*). According to this explanation therefore: (1) Verner's Law originated in Proto-Germanic as a tendency of keeping word stress on the initial syllable; for this purpose, certain adjustments had to be made in the voiceless obstruents. (2) When the Germanic immigrants arrived on the coasts of Finland and Estonia during the Bronze Age and Iron Age (ca 1400-900 BC and/or 200-400 AD), they still retained the custom of pronouncing words according to Verner's Law. The immigrants were unable to alter their pronunciation even when learning Finnish and Estonian. The result was that all Finnic languages that had contacts with the Germanic people aquired the same custom (Lauri Posti 1953).

ESTONIAN ACCENTS

The Estonian tonal opposition between Quantity 2 (Q2) and Quantity 3 (Q3) words is manifested as a late versus early tonal peak. The tone of Q2 is often referred to as "rising" or "level" and that of Q3 as "falling". From the point of view of its origin, the Estonian rising/level tone can be associated with the IE and Germanic grave tone. The late peak of the Estonian Q2 words sounds much like an additional pulse after the stressed syllable. The Estonian rising/level tone of Q2 occurs in the forms that originally had a closed second syllable (e.g. the genitive singular still had the suffix *n* as in *linnan*) while the falling tone of Q3 occurs in the forms that had an open second syllable (e.g. *linna-ta* and *linna-han*). In the Germanic languages, it was exactly this long unstressed syllable that was one of the reasons for the use of the grave tone.

The seemingly complicated tripartite durational system of Estonian is an automatic reflection of the Lappo-Finnic foot isochrony. The old foot isochrony experienced one minor change, however, when the Germanic and/or Baltic immigrants adopted Estonian: they "misunderstood" the scope of foot. Foot was originally defined as the sequence from the beginning of the first vowel to the end of the second syllable vowel; thus the foot was represented, for example, by the underlined portions of the following words. *lina, linnad, linnast(a)*. When the Germanic/Baltic people adopted Estonian, they misunderstood the scope of the foot in that they included the final consonant in the foot (they did this probably because the second syllable consonant really was relevant in consonant gradation of obstruents). Thus in their speech, foot isochrony was different from the original one: If the second syllable was long (if it ended in a consonant) the foot was "too long" and its first syllable had to be shortened: *linnad* > *liñnad*. But if the second syllable was short (if there was no consonant at the end), the first syllable had to be lengthened: *linna(ta)* > *liṅna*. The shortening and lengthening was placed on the second mora. Another reflection of the Lappo-Finnic foot isochrony is the shortening of the second syllable vowel after Q3 (e.g. *liṅnă* and *jaàmă*).

HALF-LONG VOWEL AND VOWEL BALANCE

A common feature of the languages of the Baltic Sea area is that word forms are divided into two categories, shortsyllables and longsyllables (e.g. *ma.nan* vs. *maa.nan/man.nan*) according to the structure of the first syllable. The phenomenon is based on foot isochrony, here assumed to be of Lappo-Finnic origin. It occurs in the Finnic languages and in the northern dialects of Swedish and Norwegian.

The eastern Finnish and main Estonian type with a mid-long second syllable vowel and the F0 peak on the first syllable represents the oldest type, which I assume to be "original". (b) The south-western Finnish and westernmost Estonian type with a very long second syllable vowel and the F0 peak on the second syllable vowel represents the type that was used by the Swedish immigrants of ca 1000-1200 AD. Their word prosody

was a result of a mixture of the Lappo-Finnic foot isochrony and the old IE/Germanic/Scandinavian grave tone: The second syllable with a half long vowel was interpreted to be "long" and was therefore pronounced with the grave tone.

The vowel balance of northern Scandinavia was originally a purely durational phenomenon and identical with the old Finno-Lappic foot isochrony: the second syllable vowel was relatively long after a short first syllable (e.g. *livà*) and short after a long first syllable (e.g. *bristă*.); Kock 1901, p. 91-99. Later the phenomenon became a qualitative difference of vowels because the long and short vowels developed differently (e.g. *livå* and *brista/brist*).

SCANDINAVIAN WORD TONES

Swedish and Norwegian are tone languages that have the tonal opposition of the acute and grave tones. The Scandinavian tones were already present in the phase when practically all Scandinavian words were disyllabic, and when the most relevant demarcation line between two types of words lay between the words with a long/half-long second syllable vowel and those with a short second syllable vowel. The former got the grave tone, the latter did not.

STØD

The stød that occurs in the Baltic languages, Danish and Livonian belongs, no doubt, to the prosodic peculiarities of the Baltic Sea area. I have shown elsewhere (Wiik 1989) that the Livonian stød is a remnant of an earlier syllable boundary. When a word like *va.lo*, for example, lost its final vowel because of apocope, the physical manifestations of the syllable boundary remained and began to be interpreted as the stød. By my definition, the stød is an "ex-syllable boundary", i.e. a phenomenon that has the physical manifestations of a syllable boundary, but which no longer is distributionally a real syllable boundary (as it does not occur in the natural position of a syllable boundary, immediately before a CV sequence). The same explanation holds for Danish. It is commonly maintained that the stød usually occurs in those Danish words that earlier used to be monosyllabic (these words usually have the acute tone in Swedish). However, the "old monosyllables" are exactly those words that even earlier (1) were disyllabic and (2) became monosyllables by losing their second syllable vowel. In the earliest phase, practically all words were disyllabic. A radical change took place when the second syllable long vowels were shortened and short vowels deleted; theoretically *man.naa* > *man.na* and *man.na* > *man.n*; real examples of Danish monosyllables with stød: Proto-Germanic *staina-* 'stone' and *hurna-* 'horn' > Modern Danish *ste.n* and *hor.n*.

VOWEL HARMONY AND UMLAUT

At the time of the early FU-IE contacts, the Finno-Ugrians had to pronounce words that contradicted vowel harmony. I assume the result was similar to what is happening today when Finns force themselves to pronounce a word contradicting vowel harmony. For example, they often pronounce the difficult loan word *olümpialaiset* as *ólümpiàlàiset*. Instead of clear front vowels *ü, ö,* and *ä* or clear back vowels *u, o,* and *a*, they pronounce intermediate vowels *ù, ó* and *á*. I assume the same happened when their ancestors acquired Proto-Indo-European/Proto-Germanic thousands of years ago. A simple real example from that period would be that the prehistoric Finns had to pronounce the word *handir*, which contradicted vowel harmony (the harmony classes at that time were essentially *i-e-ä* and *j-o-a*). The result was something like *hàndịr* with two vowels of intermediate quality. Later the intermediate vowels of the first syllable became independent new phonemes, as the complementary distribution of the original back vowels *u, o, a* and the new intermediate vowels *ù, ó, á* was broken by the mergers in the conditioning unstressed vowels.

SECONDARI WORD STRESS IN THE RHYTHMIC WORD STRUCTURE

I.Loginova
Russian People's Friendship University, Moscow, Russia

ABSTRACT

The subject is the rhythmic structure of polysyllables with a primary (further: PS), one or more secondary stresses (SS): the morphological secondary stress (MSS), the rhythmic one (RSS), their functions, placement in a word, ways of manifestation, interaction, in Russian, English, German, French, according to the pronunciation dictionaries, auditory samples collection and on experiments.

GENERAL PRESENTATION

In the majority of word stress languages there is only one PS in a word: in Russian different syllables may be stressed; in German it tends to occur in the stem syllable and it falls on the beginning of words.

However in English words having separable prefixes and in compound words there may be two equally strong PS, and in abbreviated compounds and abbreviations - even three or four. For instance: "week-"end, "radio"active, "normal"school, "il"legally, USA/"ju"es"ei/. The word accentual pattern with equal stresses is considered to be productive in English.

In word stress languages a polysyllable can possess not only a PS but also one or more SS of various types which differ in position and functions and ways of the phonetic realization. One of the types of the SS - the MSS - is determined by the morphological structure of the word and occurs in compounds and abbreviated compounds (more often in polysyllables), abbreviations, in the word whith accentually prominent morphemes (prefixes, suffixes).

Thus, in Russian this stress may mark a non-final stem of the compound or abbreviated compound ('лесозаго"товки, 'соб"кор), a non-final element of the initial abbreviation (ЧПИ /'че"пэ/, РФ /'эр"эф/), some prefixes either bo-rrowed or Russian by origin: ('ультрасовре"менный, 'деколониали"зация, 'послеопераци"онный).

The word of the analogous morphological strusture with two prosodic heads can be found in English (a"larm-"clock, "air-'hostess, "any'body, ABC/"eibi:"si:/, "ante'chamber, "sub'structure), in German ("Schreib'tisch, "Auf'bau, 'rot"weiß). It should be said that in German both prefixes (separable - "trennbare"), and suffixes with unreduced vowels ("schwere") may be accentually prominent morphemes ("*aus*'nehmen, "*Nach*er'zählung, "Wirt*schaft*, "Frei*heit*, "arbeit*los*); in English - "separable" prefixes and "prominent" suffixes with an unreduced vowel ("*sub*va'riety, "*non*-con'ducting, "*amphi*the'atre, "demonstr*ate*, "beauti*fy*, "social*ize*).

Depending on the number of stems and accentually prominent morphemes there may be one or several MSS: двадца'типя'тиру"блёвый, 'проф'техобразо"вание, МГУ/'эм'гэ"у/; "air-'speed'meter, "air"vice-'marchal; "Selbstbe'stimmungs'recht, "Rot'kreuz'schwester.

The mutual placement of the PS and MSS is a characteristic feature of the rhythmical structure of a word in different languages. In Russian MSS always precede a PS; in German it more often follows a PS (it corresponds to the semantic value of different parts of German compound word: in the first place there is a determining part, in the second - a determined one). Compare compound words of an analogous structure in Russian and German: само'лётостро"ение - "Flugzeugproduk'tion, 'радио"станция - "Radiosta'tion, 'трёх"атомный - "dreia'tomig.

The tendency for a strong beginning in German words determines shifting of the PS on to a separable prefix leaving only a MSS on the stem (both in a separate word and in speech, in which separable prefixes are in a final position in a phrase: "ab-'schreiben - Sie 'schreiben diesen "Text "ab).

In English the majority of compound words and a few simple words having separable prefixes follow the pattern: PS + PS, PS + SS (in which a prefix has a PS, the stem - a SS): "goal'keeper' "finger-'alphabet; see other examples above. Simple words having unseparable prefixes may be formed on the pattern: SS + PS (with a PS on the stem and a SS on the prefix): 'inter"action, a"lign - 'rea"lign; the words having "prominent" suffixes may be formed on the same pattern with the PS on the suffix a SS on the stem: em"ploy - 'employ"*ee*, "engin - 'engi"n*eer*, gre"nade - 'grena"d*ier*, "picture - 'pictu"*resque*.

Thus, in spite of the tendency of the PS to fall on the stem syllable in English and German, words having only a PS accentually prominent non-stem morphemes in a word having two stresses, often take a PS, whereas in Russian non-stem morphemes in a word having two stresses receive only a MSS.

A MSS falls on the same syllable of the stem as a PS in the parent word: ва"гон - ва'гоноре"монтный, 'ety"mology - "folk-'ety"mology, "Birke + Ge"hölz - "Birkenge'hölz, Bi"llett + "Ausgabe - Bi"llett'ausgabe. When a parent word has a stress on a final syllable, and in case of a shifting stress in a parent word it falls on the syllable which is stressed in one of the derivatives from the same root or one of the grammatical forms of the word: кисло"та - ки"слоты - ки'слоту"порный, "image - i"magin - i'magi"nation. In Russian stems with the sequence of sounds -оло-, -оро-, -ере- the MSS is shifted to the first sullable of these sequences: моло"ко - (мо"лочный) - 'молокора"зливочный; connective vowels -о-, -е- in Russian compound words are always unstressed. Accentually prominent morphemes retain the etymological placement of stress: 'около"солнечный, 'антиоб"щественный, "Wieder'sehen, "über'laufen.

Initial abbreviations (in letters) have the following stress patterns in different languages: in Russian - MSS + PS (as well as in compounds): ФПК/'эф'пэ"ка/, МГПУ/'эм'гэ'пэ"у/; in German they have one PS on the last component: FDP/efde:"pe:/; in English they have equally strong stresses on each component or on the first and the last components: M.P./"em"pi:/, BBC/"bi:"bi:"si:/ = /"bi:bi:"si:/, USA/"jues"ei/.

The way of the MSS realization is analogous to the phonetic correlates of the PS but with a lesser degree of prominence. In languages characterized by the qualitativ difference between stressed and unstressed vowels absence or presence of the qualitative reduction of a vowel becomes one of the criteria in ascertaining accordingly presence or absence of a MSS or accentual prominence in a word. Compare: горнолыжный/'горнъ"лыжный/-/гърна"лыжный/, хлебоуборочный/'хл'ебъу"бо-/ - /хл'ьбъу"бо-/, a"gronomy = 'agro"nomics, ex-"pose - 'expo"sition, "commune -

com"municate, ver"sorgen, "Arbeiterschaft.

A MSS regulates some other phonological processes in the word stem analogous to a PS in the given stem used in isolation. For instance, in Russian word having a MSS there are signals showing the end of the adjacent words, these signals are connected with devoicing of the voiced consonants, darkening of palatalized sounds, absence of palatalization in as-similation, placement of allophones /j/ on the stem ends: 'близле"жащий /с/, 'завот"делом /ф/, 'медисти"тут /-дын-/, 'мед-"техника /-тт'-/, 'стройо"тряд /-ойа-/, 'дет"ясли /-тja-/.

A MSS performs the same constitutive function to a stem within a compound word as a PS does to the whole word. In case of loss of the meaning by the stem a Russian MSS is lost, this process is accompanied by phonetic changes. In connection with this process changes in the phonetic and accentual characteristics of words are observed.

In English polysyllable having two equally strong stresses the loss of the meaning by the constituent parts of the word leads to decrease of the degree of prominence and replacement of one of the PS by a MSS. The loss of the MSS does not take place; stressed syllables (having PS or MSS) are distinctly opposed to unstressed ones. A MSS is retaining in German in which a compounding is productive way of word-formation and accentual features of the stem syllable becomes one of the means in word- identification.

Another type of the SS - a RSS- is determined by the length of the rhythmic structure of a simple or a compound word, and it occurs in a polysyllable (or a rhythmic group) containing a long prestressed sequence of syllables), it does not depend on the morphological structure of a word.

This type of stress is most vividly represented in French having no word stress in the strict meaning of this term as individual characteristics of each word form; and primary rhythmical stress falls on the last syllable of the rhythmical group. In each rhythmic group there may be one or several more SS on the odd from the end of the rhythmic group syllables: avec 'un cou"teau, il 're-vient "tard, contre la 'veri"té, Na-'bucho'dono"sor. If a primary final stress has a delimitative function of division the utterance into rhythmic groups, a RSS helps the convenience of pronouncing and it forms rhythm of French speech.

In English a RSS is manifested in different ways. It may be obser-ved in simple polysyllable words and words having unseparable prefixes before a PS (and it is different from a MSS which more often follows a PS): 'maga"zine, 'eco"nomics, 'confi"dential. A nu-mber of English suffixes receive a PS and in this case in a long prestressed sequence of syllables there may occur a RSS coinciding or not with a PS in the parent word: "refuge - 'refu"gee, ab"sent - 'absen"tee. The other suffixes (often of Greck-Latin origin) remain unstressed and regulate the placement of a PS in a word shifting it on a vowel before a suffix that is the second or the third syllable from the end. In this case in the initial prestressed sequence of two or more syllables there occurs a RSS coinciding or not with the stem syllable: "negativ - ne-"gation - 'abne"gation - 'nega"tivity, "criminal - 'crimi"nology, "lexical - 'lexi"cography.

Linguists usually do not differentiate between two types of English second stress possibly be-

cause of the identity of the means of their phonetic realization, in particular, absence of qualitative reduction of the vowel. At the same time there are views on some "prominence" of the number of suffixes containing unreduced vowels but having no SS: "celebrate, "normalize, "satisfy. Such "prominence" is similar to "prominent" ("schwere") morphemes in German (having unreduced vowels), also being considered unstressed. Since qualitative and quantitative reductions are usually interconnected, such syllables may be percieved as weakly stressed against the background of reduced unstressed syllables.

A RSS in Russian word of any morphological structure is found on the fourth or the fifth prestressed syllables; it is accompanied by a qualitative vowel reduction, and it is expressed only by prolongation and probably by strenthening of a prominent syllable: 'запатенто"вать /зъ-/, 'целесоо"бразно/цъ-/, 'революци-"онный /р'ь-/. This type of stress is never observed in post-stressed sequence of syllables despite of the length of this sequence: "сковородами /ъ/, "жаворонков /ъ/.

Russian word may possess only one RSS which excludes a MSS and may or may not coincide with it in placement in a word (when it coincides, that is there is a MSS on the fourth or fifth prestressed syllable, a RSS helps to retain a MSS in Russian word). It leads to different variants of pronunciation of a word: 'многонацио"нальный /'многъ-, 'мнъгъ-/, 'трагикоме"дийный /'тра-, 'трък-/, же'лезнодо"рожный /жы'л'е-, 'жъл'ъ-/.

Thus, in Russian the types of rhythmic patterns of the words having two or many stresses are different from the other languages. They may be represented inpatterns: MSS + PS (or MSS + MSS... + PS), RSS + PS.

FRENCH AND KOREAN PLOSIVES: A COMPARATIVE ANALYSIS

H-Z Kim and A. Bothorel
Institut de Phonétique de Strasbourg - USHS
22 rue Descartes
67084 Strasbourg Cedex, France

ABSTRACT

The aim of this study is to examine, for French and Korean, differences between acoustic temporal cues in the production of plosives by French and Korean bilinguals.

The Korean consonantal system has three unvoiced plosives :

	lenis	aspirated	glottalized
bilabials	p	p^h	p'
alveodentals	t	t^h	t'
palatals	c	c^h	c'
velars	k	k^h	k'

It is admitted for French that there are two series of plosives :

unvoiced	p	t	k (fortis)
voiced	b	d	g (lenis)

They differ as to their voicing status (voiced/unvoiced) and their degree of tension (lenis/fortis).

The main question addressed here is, how can one explain the difference in behaviour for apparently the same elements that are in a similar context? The answer to this question is discussed in relation to the notion of the phonetic context and its influences. An additional theoretical notion should also be taken into account; that of a system, a decisive factor that can not be dissociated from the notion of "phonological constraints".

A series of experiments are presented, in which identical stimuli are used to examine similar cue-trading relations in the perception of the voicing contrast in word stops in French and Korean. Predicted cross-linguistic differences are found in the basic category boundary and in the case of cue trading between VOT and aspiration.

INTRODUCTION

The native language (L1) one learns in early childhood and a second language (L2) learned later in life often influence one another. The authentic pronunciation of phones in a foreign or second language (L2) may require the establishment of new phonetic categories. Even though the acoustic differences resulting from these different articulators may be detectable (Flege and Hammond, 1983; Flege, 1990), listeners seem to classify realisations of /t/ in Spanish and English as the "same" at a phonological level. For example, Bohn and Flege (1990) found that Spanish monolinguals consistently identified long-lag English [t^h] tokens as /t/ in a two-alternative forced-choice test. English monolinguals identified Spanish short-lag [t] tokens as /t/ in the majority of instances even though they had VOT values that, in an experiment with synthetic stimuli, would be expected to give rise to the perception of /d/ (Williams, 1977; Flege and Eefting, 1986).

Second language (L2) speech production researches have shown that few late learners fully differentiate /p,t,k/ in their two languages if voiceless stops in the L1 are realized with short-lag VOT values and voiceless stops in the L2 are realized with long-lag VOT values. Previous studies have shown that many adult L2 learners produce English /p,t,k/ with significantly shorter VOT values than English monolinguals, but with significantly longer VOT values than monolingual native speakers of the learners'first language (L1) (Port and Mitled, 1983; Nathan, 1987; Flege, 1987). When late learners'VOT values for English /p,t,k/ are intermediate to the values observed for monolingual speakers of the L1 and L2, they are said to have been produced with "compromise" values (Williams, 1980). The seeming limitation on how accurately VOT in English /p,t,k/ is produced also seems to apply to adolescents and older children (Flege and Eefting, 1987). Flege and Hillenbrand (1984) hypothesized that an upper limit exists on the extent to which late L2 learners can approximate the phonetic norms of English for /p,t,k/ based on the observations that compromise VOT values.

The results of other L2 production studies, on the other hand, suggest that even early learners may fail to produce English /p,t,k/ authentically. Caramazza et al. (1974) found that native French speakers who began learning English by the age of 7 produced English /p,t,k/ with significantly shorter VOT values than native speakers of English. Flege and Eefting (1987) also found that native Spanish adults and children who began learning English as a second language by the age of 5 to 6 years produced English /p,t,k/ with significantly shorter VOT values than age-matched groups of native English subjects. These studies suggest that early learners may be unable to fully differentiate /p,t,k/ in L1 and L2, and thus, support the view that both the L1 and L2 phonetic systems remain activated to some degree.

In summary, previous research have established that late learners are apt to produce English /p,t,k/ with VOT values that are too short for English. But it remains uncertain as to whether early learners will also differ from native speakers of English, or if they will fully differentiate corresponding L1 and L2 stops. Few previous studies have examined whether learning L2 affects how bilinguals produce stops in their L1. It appears that no previous study has directly compared the production of L2 stops by early and late learners.The aim of this study is to determine if such sub categorical phonetic differences between native and non-native speakers will suffice to cue the detection of VOT (consonant duration and duration of the preceding vowel).

I. EXPERIMENT

A. Methodology

1. Subjects

Two groups of monolinguals (six males: 3 French, 3 Koreans) and one group of bilinguals (3 Koreans: males) participated as paid subjects. The native French and the native Korean monolinguals differed little in mean age (31 *vs* 25 years). The native French were students at the University of Strasbourg. The native French-speaking Koreans did not begin learning French until they were adults. The subjects in this group were native speakers of Korean who learned French as a second language. The learners indicated that they were first exposed massively to French when they started their university studies in Seoul between the ages of 18 and 19.

2. Materials and procedures

Owing to phonological differences between French and Korean, it was not possible to find a list of matched French and Korean words. Each speaker uttered a series of words in carrier sentences (with comparable syllabic structures) at a normal self-selected speaking rate. The reading task was modeled at a moderate speaking rate on the instruction tape using a list of utterances resembling those on the randomized list.

3. Measurements

Native (Group 1: Control Group) and non-native (Group 2: Experiment Speakers Group) produced minimally paired /VCV/ syllabic structures. Each of the test words occurred three times on the French and Korean lists. A total of 36 phrases in Korean and 27 phrases in French from the middle of each list were digitized at 10 kHz.

B. Results

The subjects, made identifications of medial stops as /p,t,k/ French or /p,t,k,p^h,t^h,k^h,p',t',k'/ Korean. In the experiment, VOT was measured in the French words of carrier sentences (with comparable syllabic structures) spoken by 3 French monolinguals and by 3 Korean bilinguals who had learned French and in the Korean words spoken by 3 Korean monolinguals. Each speaker uttered a series of words in carrier sentences (with comparable syllabic structures) at a normal self-selected speaking rate. The study aimed first at determining the differences between the production of the voiceless stops and the differences in the degree of palatalization (revealed by VOT) for the two languages using the VOT values, and second, at analysing the characteristics of the French consonants produced by native French-speaking Koreans.

Results discussed mean VOT values for French and Korean voiceless stops tokens that were produced in utterance-medial position. The value shown here averaged across the /a/ context. VOT, we

have noticed, is the principal cue in distinguishing the three Korean plosives, as both absolute and relative VOT values for aspirates are different from those of the non-aspirates.

In Korean, VOT is, therefore, among the indicators that permit us to distinguish the three categories of Korean voiceless stops, which is very important in making a distinction between the aspirated consonants and the glottalized consonants for utterances in medial position.

As expected, the monolingual French speakers'VOT is superior to the voiceless glottal stops of VOT of the French-speaking Koreans.

Also, as expected, the French-speaking Koreans produced the French plosives with shorter VOT values than the French monolinguals, as their VOT correspond to values comparable with those of the glottal plosives for Korean.

The duration of Korean plosives permitted us distinguish between the three categories, i.e. aspirates, glottals and lenis in an intervocalic context.

Total consonant duration in Korean allows distinguishing aspirates, from glottals and lenis consonants, when flanked by vowels; the voiced counterparts are always shorter.

With regards to the duration of the preceding vowel, we recognized that this measure is not an indicator sufficient to distinguish the voiceless plosives of French-speaking Korean, even less in the case of the aspirated plosives and of the glottalized plosives. The duration of the preceding vowel is clearly shorter before the lenis plosive velars in comparison with the lenis plosive bilabials and the lenis plosive alveodentals.

In the case of the French plosives of the native French, (as well as in the case of the Korean bilinguals), the duration of the preceding vowel, being practically identical, does not constitute an indication of differentiation.

II. FINDINGS, DISCUSSION & IMPLICATION

This study will be very useful for comparing our results with any other recent resarch on VOT especially in FLEGE's works. This experiment yielded results that were very much the same as those obtained in his experiments. Both the native French, the native Korean monolinguals and the Korean bilinguals produced French and Korean plosives

The bilingual Korean subjects had larger VOT differences in the sentence condition, where the French and Korean sentences were produced in alternation.

These changes, in the production of VOT of the bilinguals, are more important than the results obtained by CARAMAZZA for French-English bilinguals. This is also a very strong tendency, as in the research of WILLIAMS (1977), for the voiced /b/ of bilinguals in English.

CARAMAZZA & YONI-KOMSHIAN (1974) have concluded that VOT is a sufficient phonological cue for the distinction of the homorganic stop consonant of French spoken in Paris. They have also proposed an explanation for the observed differences between French and Canadian French based on a linguistic change hypothesis.

FLEGE (1988, 1990), on the other hand, hypothesized that complete separation of sounds in the L1 and L2 phonetic inventories is possible, at least for early learners.

Previous studies have shown that many adult L2 learners produce English /p,t,k/ with significantly shorter VOT values than English monolinguals, but with significantly longer VOT values than that ofmonolingual native speakers of the learners L1 (FLEGE & PROT, 1981; PORT & MITLER, 1980; HATHAN, 1987; FLEGE, 1987A; NATOR, 1987; LOWIE, 1988).

According to FLEGE & EEFTING (1987), it appears that proficient Dutch speakers of English produced Dutch /t/ with shorter VOT values than non-proficient subjects, suggesting they formed a new category for English /t/.

This finding corroborates our results. As mentioned earlier, VOT is the strongest cue in differentiating Korean plosives. VOT in French is superior to that for the glottal class in Korean. For French-speaking Koreans (stressed = 22 ms; unstressed = 17 ms), VOT values are not high and correspond to those obtained for French speakers (stressed = 25 ms; unstressed = 18 ms), as they also correspond to those measured for glottalized Korean plosives (16 ms). However, VOT for the glottals in Korean are shorter than that for French spoken by Koreans.

The findings of FLEGE (1987) presented here indicate that adults are capable of learning to produce new phones in an L2 and of modifying their previously established patterns of articulation when producing similar L2 phones. It appears that the mechanism of equivalence classification leads to identifying acoustically different phones in L1 and L2 as belonging to the same category. This may, ultimately, prevent them from producing exactly similar but new phones authentically.

CONCLUSION

We can therefore conclude that Korean bilinguals who speak French realized another form for French /p,t,k/. We are dealing here with a new category for French /p,t,k/. Because VOT values of French voiceless stops /p,t,k/ produced by Korean bilinguals are very similar to VOT values by French monolinguals. But it seems, under examination of consonant duration, that neither Korean glottalized consonants nor French voiceless stops are used by Koreans who speak French and, with regard to VOT, it is not used for Korean consonant aspirated. This may correspond to a new category of voiceless plosives (for French-speaking Koreans).

These results are interpreted to mean that individuals who learn L2 later in life are also able to establish phonetic categories for sounds in the L2 that differ acoustically from corresponding sounds in the native language. The results strongly suggest that the late L2 learners porduced /p,t,k/ with slightly longer VOT values in French than Korean glottalized plosive and shorter VOT values in French than Korean aspirated plosive by applying different realization rules to a single phonetic category.

REFERENCES

[1] FLEGE J., HAMMOND R., (1982), "Mimicry of non-distinctive phonetic differences between language varieties", *Stud. Sec. Lang. Acquis. 5*, pp. 1-18.

[2] FLEGE J., (1984), "The detection of French accent by American listeners", *J. Acoust. Soc. Am. 76*, pp. 692-707.

[3] BOHN O.S., FLEGE J., (1990), "Perceptual switching in Spanish/English bilinguals; Evidence for universal factors in voicing judgments", *J. Phon.* (submitted).

[4] FLEGE J., (1990), "The production of cognate English words by native speakers of Spanish: More evidence for the distinction between phonetic implementation and realization", submitted to *J. Phon.*

[5] WILLIAMS L. (1977), "The voicing contrast in Spanish", *Journal of Phonetics, 5*, pp. 169-184.

[6] FLEGE J., EEFTING W. (1988), "Imitation of a VOT continuum by native speakers of English and Spanish : Evidence for phonetic category formation", *Journal Acoust. Soc. Am., 83*, pp. 729-740.

[7] PORT R.F., MITLEB F.M. (1983), "Segmental features and implementation in acquisition of English by Arabic speakers", *Journal of Phonetics, 11*, pp. 219-229.

[8] NATHAN G. (1987), "On second-language acquisition of voiced stop", *Journal of Phonetics, 15*, pp. 313-322.

[9] FLEGE J. (1987), "The production of *"new"* and *"similar"* phones in a foreign language : evience for the effect of equivalence classification", *Journal of Phonetics, 15*, pp. 47-65.

[10] WILLAMS L. (1980), "Phonetic variation as a function of second-language learning", in *Child Phonology, Vol. 2 Perception*, edited by Yeni-Komshian G., Kavanagh J. & Ferguson C. (Academic, New York), pp. 185-216.

[11] FLEGE J., EEFTING W. (1987), "Cross-language switching in stop consonant production and perception by Dutch speakers of English", *Speech Commun. 6*, pp. 185-202.

[12] FLEGE J., HILLENBRAND J. (1984), "Limits on pronunciation accuracy in adult foreign language speech production", *J. Acoust. Soc. Am. 76*, pp. 708-721.

[13] CARAMAZZA A., YENI-KOMSHIAN G.H. (1974), "Voice onset time in two French dialects", *Journal of Phonetics, 2*, pp. 239-245.

THE LANGUAGE OF THE WEST SIBERIAN MENNONITES

T. de Graaf and R. Nieuweboer
Department of Linguistics,
Groningen University

ABSTRACT

Plautdiitsch, the language used by Mennonites in many parts of the world, is a descendant of West Prussian Low German dialects. Many of its peculiarities can be explained by the two centuries of isolation from other German dialects and by contacts with other languages. Until recently, the dialect in West Siberia could be studied by Soviet scholars only, but in the last few years it has become possible also for others to do ethno-linguistic field work in this area. The Universities of Groningen, Oldenburg and Novosibirsk study this particular variety of the Plautdiitsch language in a joint research project [1].

HISTORICAL BACKGROUND

In the beginning of the 16th century, growing discontent with the catholic church led to the foundation of a number of new religious movements. In the Netherlands the former catholic priest Menno Simons (from a small village in Friesland) gathered a number of people around him who came from various parts of the Netherlands, but also from Germany and Switzerland. They moved from the Low Countries to the Weichsel delta area near Danzig (Gdańsk in present-day Poland).

The settlers were not a homogeneous group, they spoke different languages and dialects: Frisian, Low Franconian and Low Saxonian dialects. In their new country, they lived among people who spoke various Low German dialects, which must have sounded rather familiar to them. Dutch was to be preserved as the language used in church for over two centuries, and religious literature for the Mennonites was printed in the Netherlands, but for everyday communication the dialects of the area were soon adapted [2].

The Polish state did not interfere much with the lives of the emigrants, and until the first Polish Partition (1772) the Mennonites were allowed to live according to their principles. When, as a result of this Partition, the area around Danzig became a part of the state of Prussia, the situation deteriorated significantly. In 1789, a first group of settlers set off for Southern Russia, to areas later to become part of the Ukraine. In 1803/4 a second group of Mennonites left the Weichsel area for Southern Russia. These two groups of settlers came from different parts of the Weichsel delta area and different social backgrounds and they spoke somewhat different dialects. The two dialects evolving in the new environment in the two colonies, Chortitsa and Molochna, both being mixtures of Western Prussian speech varieties, reflected the differences in geographical, social and historical background of these two groups [3].

A major setback for the Mennonites was the abolition of their privileges in Russia in the 1870's, which resulted in a large emigration to Northern America (Canada and the United States) [4]. The founding of new colonies continued, however, and after 1910, a group of settlers left for Siberia and founded colonies in the Orenburg region and the Kulunda Steppe near the border with Kazakhstan.

After the October Revolution, the situation for the Russian Mennonites became very difficult. In the 30's, the first mass deportation took place, and when World War II started, all ethnic Germans in the Soviet Union faced labour camps and, again, deportation. The situation improved somewhat after 1953, but it lasted many years before emigration again became a possible alternative. The colonies in the Ukraine had disappeared, and most Mennonites now lived in Siberia and Kazakhstan. Only in a few areas in South Western Siberia Mennonites still lived in ethnically homogenous villages, in most other parts of the country they were scattered amongst many other nationalities. In the villages in the Altai Region we visited, only a few years ago almost 100% of the population were Mennonites, but now in some villages they are a minority. During our stay in the Altai region, we studied the complicated language situation, in particular its phonetic/phonological aspects. We collected many data on language use by the local inhabitants and found interesting cases of code switching and interference.

THE PLAUTDIITSCH LANGUAGE

The Plautdiitsch language as it is used today in Mennonite communities all over the world is the descendant of West Prussian varieties of Low German. The two century isolation in a non-German speaking environment has resulted not only in a considerable amount of loanwords from the surrounding languages, but also in a somewhat different and partly accelerated development of a few elements already present in the Weichsel delta dialects. The resemblances between Plautdiitsch and Dutch, or rather the Low Saxonian dialects of the Dutch language, were sometimes used to indicate the non-German character of Plautdiitsch and thus to 'proof' the Dutch origin of its speakers [2].

In most Mennonite communities two different varieties of the language are used, based on the Old Colony or Chortitsa dialect, and the New Colony or Molochna dialect respectively. The variety spoken in the Altai Region is an example of a mixture of these two dialects. The following list shows some of the most typical differences between the dialects of Chortitsa and Molochna, and the forms used in the Altai Region, which we investigated during our field work in 1992 and 1993 [5, 6].

Khortitsa	Molochna	Altai	
[k'ɪk'ən]	[t'ɪt'ə]	[t'ɪt'ə]	*to look*
[lɪg'ən]	[lɪd'ə]	[lɪd'ə]	*to lie*
[hy:s]	[hu:s]	[hy:s]	*house*
[by:ən]	[bu:ə]	[by:ə]	*to build*
[ku:lən]	[ko:lə]	[ko:lə, ku:lə]	*coal*
[møɔkən]	[mɔɐkə]	[moɐkə]	*make, do*
[zeɪt]	[zeɪt, zəɪt]	[zəɪt, zɔɪt]	*sweet*
[plaut]	[plɔ:t]	[plɔ:t]	*flat*
[ji:]	[zeɪ, zəɪ]	[zəɪ, zɔɪ, ji:]	*you (polite form)*

The consonant systems seem to be practically identical, with the exception of the development of the palatalized phonemes originating from /k, g/ before or after front vowels: the Chortitsa dialect has [k', g'], the Molochna dialect has [t', d']. The main differences are found in the vowel systems: most long vowels and diphtongs have separate realizations in the two varieties. Within the Altai dialect, great variation in the vowel system is possible, so that the actual pronunciation of a word may differ from speaker to speaker and from occasion to occasion.

Two of the differences bear resemblance to those found in Dutch dialects: standard [œy] as in *huis* (*house*) corresponds to [u] (the older form) in some varieties of Low Saxonian, and [y] (a later development) in others; the infinitive endings [(ə)n] and [ə] are found in the Low Saxonian and the Low Franconian dialects respectively.

As we have seen in the above, Plautdiitsch has a striking peculiarity: a number of palatalized consonant phonemes. In a few dialects in or near the Weichsel delta area, /k/ in front of or following a palatal vowel was realized as [k'], [t'] or [tʃ], and in Plautdiitsch this development later continued, resulting in the three new phonemes /k'/ or /t'/, /g'/ or /d'/, and /n'/. In West Siberian Plautdiitsch, original /k/ has become /t'/ in the following positions:

1. in front of palatal vowels:
[t'oɐʃ] High German *Kirsche*;
[t'a:rps] *Kürbis*
2. in front of palatal vowels, before [l,n,r,v]:
[t'ləɪdʌ] *Kleider*; [t'li:n] *klein*;
[t'nepəl] *Knüppel*; [t'na:ls] *Cornelius*;
[t'rɪç] *Krieg*; [t'rɪpst] *kriechst* (from the infinitive [kry:pə] *kriechen*);
3. after palatal vowels:
[t'eʌt'] *Küche*; [et'] *ich*
4. after palatal vowels + /l,n,r/:
[ma:lt'] *Milch*; [drɪn't'] *trink*;
[boʌt'] *Birke*

It is believed that the palatalization may be related to the influence from Frisian. In this language, however, palatalization has a much more limited range and is restricted to old /k/ in front of palatal vowels where it has changed to [tʃ] (written as <tsj>), or, more seldom, [ts]: *tsjerke*, High German *Kirche*; *tsjettel* - *Kessel*; *tsiis* - *Käse*.

Most probably, the palatalization in Plautdiitsch has not arisen as a direct result of influence from Frisian. It is likely that some of the Low German dialects from which Plautdiitsch evolved, owed their palatalization to Frisian settlers (who moved from Eastern Friesland to the Danzig region in the Middle Ages), and that in Plautdiitsch this process developed further.

In general, Plautdiitsch shares many of the elements that distinguish the Low German dialects from High German. The sound changes [p] > [pf], [t] > [ts] that characterize the Southern German dialects have not occurred in the North, e.g.: High German (HD) *Apfel*, Dutch (D) *appel*, Plautdiitsch (PD) [a:pəl]; HD *Zeit*, D *tijd*, PD [ti:t]. Low German also has a great number of words that are unknown in High German, but not in Dutch. The Plautdiitsch word [fəndoɐɣ] (other Low German dialects have similar forms) *today*, is *vandaag* in Dutch, but *heute* in High German. These two elements give Plautdiitsch a very familiar ring to the Dutch ear. At the lexical level, the language of the Siberian Mennonites mirrors the history of these people. Most words are known from various Low German dialects: [hy:s] *house*, [drɔk] *busy*, [maŋk] *between*.

Most of the hundred or so loanwords from Dutch (Tolksdorf 1990) found in dialects spoken in the Weichsel delta area have disappeared from Siberian Plautdiitsch. A few survivors are [o:lba:səm] from Dutch *aalbes* 'black currant', [mɔ:] or [me:v] from D *mouw* 'sleeve', [ta:xəntɪç] with initial [t] as in D *tachtig* 'eighty'. The word [pi:nɪç] is the descendant of *pijnlijk* (obsolete in this meaning) 'diligent' - in Plautdiitsch is has come to mean 'quick'.

Frisian seems not to have left many traces in the Plautdiitsch lexicon, but with certainty of Frisian origin is [t'a:st] 'wedding', from Frisian *kest* 'choice'. [ʃvi:nt'ət'a:st] 'the slaughtering of a pig' must originally have meant 'the choosing of a pig to be slaughtered'.

More dominating in the lexicon are the many loanwords of Slavonic (Polish, Ukrainian, Russian) origin. In particular, the influence of Russian on all levels has become very strong, as our field work recordings show.

CONCLUSION

In the Germanic language family, Plautdiitsch claims a special place. Its long isolation from other German dialects and its close contacts with other languages have given it a specific character, which to some extent can be compared to that of Yiddish. The Plautdiitsch language, the sole descendant from the many West Prussian Low German dialects once used in the Weichsel delta area, is now spoken by Mennonites in many countries and has partly taken over the religious factor as the main identity marker for this ethnic groupe. It is a pity that a language, that managed to survive centuries of isolation and many years of prohibition, should now disappear where it has long had its most speakers – in Siberia. The increasing emigration to Germany has left many Mennonite villages russified more than decades of Soviet russification policy could accomplish. The Plautdiitsch speakers who choose to stay find it more and more difficult to provide their children with a Plautdiitsch speaking environment, and in the long run it must be feared that the language will lose much ground to Russian. In Germany, the children of Russian Mennonite immigrants will almost certainly only have passive knowledge of Plautdiitsch.

One can only hope that the language will survive in North America and the isolated colonies in South America's, where a revival can be observed.

REFERENCES

[1] Nieuweboer, R. and De Graaf, T. (1994), *The Language of the West Siberian Mennonites,*
Internationalt tidsskrift for sprog og kommunikation, 1 (1994), 47-61.
[2] Unruh, B.H. (1955), *Die niederländisch-niederdeutschen Hintergründe der mennonitischen Ostwanderungen.*
Im Selbstverlag.
[3] Quiring, J. (1928), *Die Mundart von Chortitza in Südrussland.*
München.
[4] Thiessen, J. (1963), *Studien zum Wortschatz der kanadischen Mennoniten*, Marburg: N.G. Elwert Verlag.
[5] Jedig, H.H. (1966), *Laut- und Formbestand der niederdeutschen Mundart des Altai-Gebietes,*
Berlin: Akademie-Verlag.
[6] Wall, M. und Kanakin, I.A (1994), *Das Plautdietsch in Westsibirien.*
Groningen: Lingua Mennonitica.

THE INFLUENCE OF SPEECH INTELLIGIBILITY ON THE USE OF ACCENTUATION AND GIVEN/NEW INFORMATION IN SPEECH PROCESSING

Wilma van Donselaar
Max-Planck-Institute for Psycholinguistics,
Nijmegen, The Netherlands

ABSTRACT

Two experiments were carried out to investigate whether the quality of speech intelligibility influences the use of the interdependence between ('given/new') information and accentuation in speech processing. Both normal-hearing and hearing-impaired listeners were tested. The results of the two experiments showed that, as speech becomes less intelligible, listeners depend increasingly on the interdependence between information value and accentuation.

INTRODUCTION

Previous research on speech perception has shown that the distribution of ('given-new') information and accentuation over sentences is strongly connected. Terken and Nooteboom [1] found in a series of sentence verification experiments that 'new' information is usually processed faster if it is accented, while 'given' information is generally processed faster if it is de-accented

The aim of the present research is to deepen our insight into factors influencing listeners' use of information and accentuation. Do listeners' strategies shift under the influence of, for instance, speech intelligibility? It is likely that listeners make more optimal use of the information-accentuation correspondence if word processing is complicated by the fact that the quality of speech intelligibility is low. When the speech signal is degraded, supra-segmental information is usually preserved better than segmental information, and therefore prosodic cues may be used by listeners to interpret the signal. As 'new' information is usually accented, a sentence accent may be interpreted as a marker of 'new' information. The absence of an accent on a focused constituent may be interpreted as an indication of 'given' information. The interdependence also seems useful to hearing-impaired listeners who cannot distinguish all the segments, but can perceive the intonation of utterances. However, research by Vingerling [2] led to the conclusion that the hearing-impaired subjects use speech intensity, rather than intonation, as a cue to accentuation. Linguistic patterns concerning accentuation seemed of little importance to the hearing-impaired listeners.

In the study presently described, it is assumed that, when segmental information is not easily available, both hearing-impaired and normal-hearing listeners will exploit the interdependence between information value and accentuation to the fullest, by regarding accented words as 'new' and unaccented words as 'given'.

In order to make a fair comparison between normal-hearing and hearing-impaired subjects, a two-choice task was designed. Target words were embedded in sentences and provided either 'given' or new' information, they were either accented or unaccented. The subjects had to choose between two word candidates that differed in the last consonant by one phonetic feature (e.g., mat/map).

A pretest was carried out to determine the intelligibility of the sentence materials for the hearing-impaired subjects, given a certain sound level. The intention was to achieve comparable intelligibility scores for both groups of listeners. For normal-hearing subjects USASI noise was used to reduce the quality of speech. Given a signal level of about 60 dB and a signal-to-noise ratio of roughly 0 dB for normal-hearing listeners, the average intelligibility score on test words for both groups of listeners in the pretest was 41%. A S/N ratio of 0 dB was therefore employed again in the main experiment.

An on-line two-choice task results in two different types of dependent variable: response latency (in ms) and accuracy rate (in %). It was predicted that both normal-hearing (NH) and hearing-impaired (HI) listeners would make more correct decisions and have shorter latencies for accented 'new' and unaccented 'given' target words than for unaccented 'new' and accented 'given' words.

METHOD

Material

Word material. Thirty pairs of Dutch word candidates consisting of high-frequency monomorphemic nouns that differed only in the final consonant (e.g., map/mat= file/mat) were selected.

Sentence material. The target words were embedded in sentence contexts that did not bias subjects toward one of the words of a pair. The sentence materials consisted of questions and answers in pairs. When a question contained a target word, this word was considered 'given' in the answer, otherwise it was new. The target words in the answers were either accented or not. Accent patterns were only considered correct when a 'new' word was accented, and a 'given' word unaccented. There is a difference, however, between correctness of accent pattern and correctness of response. A response was considered correct when the word was chosen that was actually offered in the answer, independent of the correctness of the accent pattern. An English transliteration of word candidates, sentence materials, and correct and incorrect accent patterns is given in Table 1.

Table 1. English example of materials. Accented words are capitalized, target words in bold. Two-choice candidates are: mouth / mouse. (N=new, G=given, A=accented, U=unaccented).

Correct accent pattern:	
N+A	Did the little girl hurt her MOUTH? She accidentally hurt her **MOUSE.**
G+U	Did the little girl hurt her MOUTH? She ACCIDENTALLY hurt her **mouth.**
Incorrect accent pattern:	
N+U	Did the little girl hurt her MOUTH? She ACCIDENTALLY hurt her **mouse.**
G+A	Did the little girl hurt her MOUTH? She accidentally hurt her **MOUTH.**

Realisation. All sentences were read by a male phonetician and recorded on DAT. Sentence accents were realised as so-called 'pointed hats'. Sentences were digitized with a sampling frequency of 10 kHz.

Subjects

Fourteen subjects, between 22 and 30 years old, who had participated in the pretest were tested. Seven subjects had self-reported normal hearing. Seven subjects had a bilateral sensorineural prelingual hearing loss. According to their audiograms, average audiometric threshold at octave frequencies form 250 Hz to 2000 Hz were 65,60,69 and 68 dB HL (mean SD=20.8). The HI subjects performed the tests without hearing aids.

Procedure

Subjects had to listen to sequences of questions and answers and simultaneously look at a computer screen. The questions were first shown orthographically and then presented auditorily. The answers were presented auditorily only. During the answers, two similar words appeared on the screen immediately after the target words (e.g., mouth-mouse). Subjects had to decide as fast as possible which of the two words they had just heard in the answer, and push a corresponding button. Reaction times were

measured from the offsets of the target words. The sessions took approximately 20 minutes.

Design

Fixed factors were Information ('given/-new'), Accent (plus/minus), and Listener group (HI/NH); random factors were Item and Subject (nested within Listener group). The percentages and latencies were subjected to separate analyses of variance, with subjects (F_1) and items (F_2) as random factors respectively.

RESULTS AND DISCUSSION

Both NH and HI subjects failed to respond on approximately 2% of the trials offered. Of the remaining responses, the HI subjects had 67% correct, and the NH subjects 85%. This response accuracy was high as compared to the pretest (41%) and probably due to the fact that a forced binary choice was employed in the main experiment, whereas free report was used in the pretest. Figure 1 gives the percentages as a function of Accent, Information and Listener group.

The effects of Accent and Listener group were significant in the analyses (at p<.001). The two-way interaction between Accent and Listener group was also significant in subject and item analyses (p<.05). The three-way interaction between Accent, Information and Listener group reached significance as well (p<.05).

NH and HI listeners followed different strategies in making their choices. Only 'accent' played a role in the decisions of NH subjects, it had a highly significant effect in a separate analysis on the NH group of listeners (p<.001). There was no interaction between Information and Accent. In the analyses on the HI group of listeners, both Accent and the two-way interaction between Accent and Information reached significance (p<.05). The latencies revealed similar response patterns, but also showed an effect of Information. Both HI and NH listeners responded faster to 'new' than to 'given' words. There was an effect of Listener group since the response times of NH and HI subjects differed 200 ms on average.

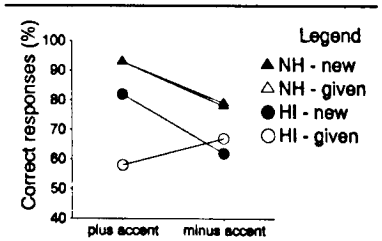


Fig.1. Correct responses (in %).

On the basis of response accuracy and latency results, the experimental predictions appear correct only for HI listeners. The question arises whether this might be due to a higher intelligibility of speech for the NH subjects. Both percentages of correct responses and latencies indicate that the noise level determined in the pretest was not optimal for comparing the two listener groups in the main experiment. In order to find out whether NH subjects would behave like HI subjects when the intelligibility of speech was more severely reduced, a second experiment was carried out.

EXPERIMENT II

In this experiment, a different group of seven NH subjects was tested with the same stimulus material and a reduced S/N ratio (of approximately -9dB).

Results

In the replication experiment, NH subjects failed to respond on 2% of the trials. Only 70% of the remaining answers was correct. This percentage closely approximated that of the HI subjects earlier. Figure 2 gives the average percentages as a function of Accent and Information. Apart from a significant main effect of Accent (p<.001), a significant interaction between Information and Accent was found in the analyses (p<.05). The latencies for NH subjects in this experiment were longer (+100ms on average) than in the previous experiment and revealed similar significancies. The outcome shows that NH listeners also use the information-accentuation interdependence when the segmental quality of speech is severely reduced.

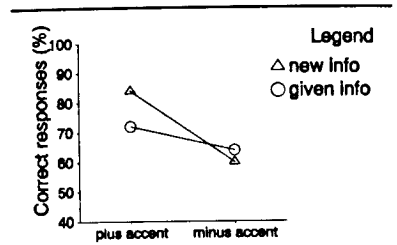


Fig.2. Correct responses NH (in %).

GENERAL DISCUSSION

A first conlusion from this study is that normal-hearing listeners make better use of the interdependence between information and accentuation when the segmental quality of speech is more reduced. This interdependence showed an asymmetry, however. Listeners did benefit from the presence of sentence accents on 'new' words, but the importance of de-accentuation for 'given' information was less clear. This asymmetry may have been induced by the experimental set-up and materials, but the phenomenon was also witnessed in experiments by others (e.g., [3]). A recent study [4] showed that de-accentuation seems less important in processing identically repeated 'given' information than in processing more implicit 'given' information.

A second conclusion is that the prelingual hearing-impaired subjects in this experiments effectively used the interdependence between information value and accentuation. From a comparison between the hearing-impaired subjects in the first experiment and the normal-hearing in the second experiment, it can be concluded that these listeners do not intrinsically differ in their use of the interdependence.

The findings in these experiments were also interpreted in terms of a temporal perspective on speech processing [4]. If listeners cannot identify word forms on the basis of segmental information only, word form identification is delayed. Incoming perceptual information on accents (e.g., intonation) and higher order knowledge on the distribution of information and accentuation are then employed to select a word candidate.

ACKNOWLEDGEMENT

This research was carried out at the Institute for Language and Speech of Utrecht University, The Netherlands. The author wishes to thank Jurgen Lentz for his help in carrying out the experiments and Sieb Nooteboom for his useful discussion of the results.

REFERENCES

[1] Terken, J.M.B. and Nooteboom, S.G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for 'given' and 'new' information. *Language and Cognitive Processes*, 2, 145-163.

[2] Vingerling, M. (1983). The perception of sentence accent in the perspective of speech processing. In M.P.R. van den Broecke, V. van Heuven, and W. Zonneveld (Eds.), *Sound Structures: Studies for Antonie Cohen* (pp. 271-280).Dordrecht: Foris.

[3] Nooteboom, S.G. and Kruyt, J.G. (1987). Accent, focus distribution and the perceived distribution of Given and New information: An experiment. *JASA*, 82, 1512-1524.

[4] Donselaar, W.A. van (1995). Effects of accentuation and given/new information on word processing. Unpublished doctoral dissertation. University of Utrecht, The Netherlands.

THE CONTROL OF INTELLIGIBILITY IN RUNNING SPEECH

E. G. Bard[1], C. Sotillo[1], A. H. Anderson[2], G. Doherty-Sneddon[2], and A. Newlands[2]

Human Communication Research Centre,
[1]Dept. of Linguistics, Edinburgh University, [2]Dept. of Psychology, Glasgow University

ABSTRACT

Speakers pronounce words less clearly when their referents are Given. Speakers' control of intelligibility is shown to reflect a rough, egocentric account of the shared Given information in dialogue. The relative reduction in intelligibility from a first to a second mention of an entity [1] was unaffected by the identity of the original mentioner, the visibility of the entity, and even by the identity of the listener.

INTRODUCTION

Although printed tokens of a word remain uniform, spoken instances will vary, even when they are produced by a single speaker in a single conversation. Part of the typical variability of speech is informative. It is created by the speaker for the listener's sake.

The duration and the intelligibility of different tokens of a word have been shown to be affected by the availability of information other than the token's acoustic shape which might aid word recognition. For example, the more predictable a word is from the sentence context in which it is read, the less intelligible it is (i.e., the lower the proportion of listeners recognizing it) when it is excerpted from that context [2]. More interesting from the point of view of the comprehension of extended discourse, word tokens are less intelligible when they refer to Given entities, both those which are, in Prince's [3] terms, *textually evoked* by previous literal mention [1, 4] and those which are *situationally evoked* by the visible presence of named item [5].

The tendency to degrade redundant tokens seems wonderfully cooperative in Grice's sense of the term[6]: speakers seem to follow a maxim of articulatory quantity in adjusting acoustic information to meet listeners' needs. In the appropriate contexts, less intelligible repeated tokens are actually helpful to listeners, for they make better prompts to earlier discourse material, either because they signal listeners to associate the word's meaning with some entity already established in a discourse model [1, 7] or simply because stored information must be called into play for successful on-line recognition of such items [8].

Unfortunately, degraded tokens are not restricted to contexts in which the listener can recover the conditioning information. Excerpted word intelligibility is equally closely correlated to predictability from sentence context (as assessed by adults) in spontaneous speech to small children and to adults [10], despite the fact that small children have far less mastery of the syntax and vocabulary of those sentences. Word intelligibility reduces across repetitions of a parent's utterance to his/her child, despite the fact that adults repeat themselves to children precisely because children appear not to have noticed earlier tokens of the utterance [5]. Word intelligibility also reduces when adults edit their recorded dictations in the workplace, even though the audio-typist will never hear the original version of the utterance because the speaker has just intentionally erased it [4].

We report several studies designed to determine how far speakers' adjustment of intelligibility is keyed to shared knowledge and how far it uses the speaker's own knowledge to model what is Given for the listener.

EXPERIMENT 1

The first experiment asked how cooperatively speakers interpret the notion 'textually evoked' in dialogue. If speakers adjust their clarity in response to their own contributions, only entities they have introduced themselves are Given and only *self*-repetitions should show a repetition effect, a loss of intelligibility from first to second (co-referential) mention. If speakers contribute to common Given set, however, then both *self*-repetitions and *other*-repetitions should show the effect.

Method

Corpus. To determine what controls the speaker's adjustments we must have an accurate idea of the information which speaker and listener command jointly and individually, including information which each has about what the other knows. For this reason all experiments reported here used word tokens from the HCRC Map Task Corpus [10]. In the 128 unscripted conversations of this corpus, pairs of speakers collaborated to reproduce on one's schematic map a route printed on the other's. Neither speaker could see the other's map. All information relevant to the task appeared on the maps. The speakers' maps differed in to some extent in the names, number, and location of landmarks. Speakers were warned in advance that their maps would not match exactly. Each speaker was Instruction Giver twice for the same map, each time with a different Instruction Follower, though no participant was ever Follower on a particular map more than once. After participating in a series of dialogues, each speaker read a list of landmark names covering the maps just used. (For other details of design see [10].) All materials were recorded on DAT (Sony DTC1000ES) using one Shure SM10A close-talking microphone and one DAT channel per participant.

These design factors make it possible to find word tokens in spontaneous speech which are supported to different degrees by knowledge shared between speaker and listener and to compare their intelligibility with 'citation' or list forms of the same items uttered by the same speaker.

Design and Procedure. The materials were all single word tokens: 48 introductory mentions of shared landmarks, that is landmarks appearing on both maps, second mentions of these by the same speaker, 48 more first mentions by one speaker and their repetitions by the other, as well the same items read by the same speakers in a list.

As in all the other experiments reported below, words were excerpted from digitally recorded materials by examining spectrogram and time-amplitude waveform representations and listening to the results of excerptions. Cut points were set at 0-crossings. Each original word speech file was multiplied, sample by sample, by a 16KHz file of random noise (where all sample values were in the range 0.5 to 1.5) of the same length. In each resulting stimulus, amplitude was related to that of the original speech and data points had the same signs as the sampled data values they replaced. Stimuli were presented from DAT with an ISI of 8 seconds. Word tokens were allocated to different presentation tapes according to a Latin square design.

Eight groups of 10 undergraduate subjects from the same population as the original speakers attempted to identify the words, with only one token of every type heard by any one subject.

Results

Scores for correct recognition were then submitted to ANOVAs by subjects and by materials. ANOVAs were performed both on *raw intelligibility*, the proportion of correct identifications, and *intelligibility loss*, the difference in rate of correct identifications between careful citation tokens and spontaneous mentions. The loss analyses remove differences in baseline citation form intelligibility from consideration. The two kind of analysis conform on the critical results. Table 1 and the reported statistics are based on intelligibility loss.

As Table 1 showed, the effect of repetition, though significant [F_1 (1, 72) = 5.90, $p < .02$; F_2 (1, 80) = 3.26, $p = .075$] was not sensitive to the identity of the original introducer of the entity [Repeater x mention: $F_1 < 1$; $F_2 < 1$, with both first mentions at .15 less than their respective citations forms and both second mentions at .23 less]. Since intelligibility is lost to the same degree for entities which either speaker has introduced, speakers would seem to retain a single common record of textually evoked given entities.

EXPERIMENT 2

The next pair of experiments asked about the applicability of the notion 'situationally evoked', that is, Given by virtue of being present to the senses as mention is made. A critical part of the map task is finding the landmarks mentioned by the other speaker. A cooperative speaker might mitigate intelligibility loss in a second mention if s/he knew that a landmark did not appear on the listener's map. The critical comparison used tokens of items which were textually evoked for both participants, because they had been mentioned before, and situationally evoked for the speaker, who could see them. Only some of these were reported as being present on the listener's map.

Table 1. Difference between intelligibility of citation and running speech forms of words repeated under several conditions (Experiments 1-3)

Stimulus Categories	Mention	
	Token 1	Token 2
Expt 1: Speakers		
Same	.15	.23
Different	.15	.23
Expt 2: Listener can see referent		
Apparently	.24	.16
Apparently not	.13	.21
Expt 3: Speaker can see referent		
Yes	.15	.30
No	.23	.42
Expt 4: Different listeners		
1st pass	.07	
2nd pass	.18	

Method

Sixty first mentions, second mentions, and citation forms were found which named unshared features in two conditions: *Apparently unshared* items were re-iterated after the listener had explicitly denied having the feature; *Apparently shared* items were repeated after the listener erroneously failed to report their absence on his or her map. All 360 word tokens were overlaid with noise and distributed among 6 groups of 9 Subjects each for identification.

Results

Clarity was lost to the same degree when the listener apparently shared the landmark (token 1: .24 less that citation form, token 2 .16 less) and when s/he apparently did not (token 1 .13 less, token 2 .21 less). Though there was an interaction between sharing and mention by subjects [$F_1(1, 48) = 9.05$, $p < .005$; $F_2(1, 96) = 2.32$, n.s.] to which we will return, second tokens did not differ by post hoc tests.

EXPERIMENT 3

Speakers' might have been insensitive to listeners' ability to see landmarks because they lacked interest in what listeners could bring to the task of recognizing words. Alternatively, visual access might affect intelligibility only in creating Given status [5], not in reinforcing it. To test this hypothesis, effects of the *speaker's* own visual access to the named item were examined.

Method

All items were items introduced and repeated by different participants, so that comparison required examing first, second, and respective citation form mentions. For 48 such sets, the repeater did not have the relevant landmark on his/her own map. For another 48, s/he did. Four groups of 9 Subjects were used.

Results

Intelligibility was reduced in one speaker's repetion of another's introduction to the same degree whether (.15) or not (.19) the repeater had visual access to the landmark named. [Mention: Min $F'(1, 116) = 10.52$, $p < .005$; Mention X visual access: $F_1 < 1$; $F_2 < 1$]. Apparently, once an item is textually evoked, neither speaker's access to supplementary visual information will affect delivery.

EXPERIMENT 4

In this experiment we ask whether the set of Given entities are marked with the identity of the individuals who know they are Given. In a map task dialogue, the Giver's strategy ought to be keyed to how much the Follower knows. Givers instruct two different Followers in a single map route. A cooperative Giver should introduce each landmark to the second Follower as clearly as s/he did to the first.

Method

The stimuli here were introductory tokens of the same landmark names uttered by the same speaker in 2 dialogues using the same map but differing in the identity of the listener. Forty-eight item triples were used (first mention to first listener, first mention to second listener, citation form) were used. Because there were some lexical duplications, the 48 were divided into 2 sub-sets and distributed by Latin square among 3 groups of 9 Subjects for identification.

Results

Second-pass introduction (textually evoked for the speaker though New for the new listener) showed greater intelligibility loss vis-a-vis the citation form (.18) than the first-pass introduction (.07) (New for both speaker and listener)(Scheffé at $p < .01$). Once an entity is entered in a representation of material textually-evoked-by-anyone, speakers appear to be insensitive to whether the current listener was witness to a previous mention.

CONCLUSIONS

The general conclusion is that intelligibility is closely controlled by the absence or presence of the named entity in a record of material textually evoked within a dialogue. Once represented in this record, an entity is named by more degraded word tokens, regardless of any other speaker or listener knowledge about the earlier mention or the entity. This arrangement bespeaks a limitation in speakers: although the basic modelling keeps a record of the shared dialogue, modeling listeners minutely while giving accurate instructions may be too burdensome a combination of tasks. In most natural dialogues, where speaker and listener are together in time and space, and where the speaker is not iterating the same message for a succession of listeners, the simplifying assumptions are correct: speaker and listener hear and see the same things, and should remember the same things about a conversation. Hence tracking any differences in situational or textual context is unnecessary. Like those ground-dwelling birds which retrieve the [11] largest visible round object when their eggs roll off the nest, speakers demonstrate what is usually a harmless oversimplification. If, however, there are large round stones near the bird, or if the listener does not share a viewpoint with the speaker, then the error is not harmless at all. The listener may be at as great a disadvantage as the oyster-catcher's egg,

In support this final claim, we can cite subsidiary results from Experiment 2. We had examined repetitions of names of landmarks which appeared on the speaker's map but not on the listener's. Now we looked at the first mentions which preceded the two responses open to the listener, correctly denying having the landmark, and incorrectly failing to report its absence. The first tokens with faulty replies were unusually unintelligible for introductory mentions. Since less intelligible tokens may be interpreted as referring to Given information, we may be dealing here with speakers' egocentric errors that carried a cost.

This work was supported by the ESRC(UK) via the HCRC. Dr. Doherty-Sneddon is now at the Department of Psychology, University of Stirling. *Address for correspondence: E. G. Bard, HCRC, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LL, UK*

REFERENCES

[1] Fowler, C. & Housum, J. (1987), "Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction", *JML, 26*, 489-504.

[2] Lieberman, P. (1963), "Some effects of the semantic and grammatical context on the production and perception of speech", *LangandSpeech, 6*, 172-5.

[3] Prince, E. (1981), "Toward a taxonomy of given-new information", In P. Cole (Ed.), *Radical Pragmatics*, pp. 223-255. New York: Acadmic Press.

[4] Bard, E., Lowe, A., & Altmann, G. (1989), "The effects of repetition on words in recorded dictations", *Proc. EUROSPEECH '89, 2*, 573-6.

[5] Bard, E.G. & Anderson, A. H. (1994), "The unintelligibility of speech to children: Effects of referent availability." *JChLang, 21*, 623-48.

[6] Grice, H.P. (1975). "Logic and conversation." In P. Cole & J. Morgan (Eds), *(Syntax and Semantics: Vol 3, Speech Acts*, pp. 41-58. New York: Academic Press.

[7] Terken, J., and Nooteboom, S. (1987). "Opposite effects of accentuation adn deaccentuation on verification latencies for Given and New information." *LCP* 2, 145-63.

[8] Bard, E. G., Cooper, L., Kowtko, J., and Brew, C. (1991), *Psycholinguistic Studies on the Incremental Recognition of Speech: A Revised and Extended Introduction to the Messy and the Sticky.* ESPRIT Basic Research Action BR3175, DYANA: Dynamic Interpretation of Natural Language, Deliverable R1.3B/C.

[9] Anderson, A. H., et al. (1991). "The H.C.R.C. Map Task Corpus." *LangandSpeech, 34*, 351-366.

[10] Bard, E., & Anderson, A. (1983), "The unintelligibility of speech to children", *JChLang, 10*, 265-92.

[11] Tinbergen, N. (1951). *The Study of Instinct. London: OUP*

MISSING DISFLUENCIES

Robin J. Lickley
Dept. of Linguistics, University of Edinburgh

ABSTRACT

Everyday experience suggests that many disfluencies pass unnoticed by listeners attending to speech. This paper presents the results of a perception experiment on a corpus of spontaneous Dutch speech, where the subjects are asked to detect disfluencies as they compare a transcript with the recording they are hearing. The results show that many disfluencies are missed by listeners even when they are trying to spot them.

INTRODUCTION

Spontaneous speech contains frequent occurrences of filled pauses (*uh, um*), repetitions (*the the door is on the left*) and false starts (*put the move the plant to the left*), all of which may be referred to as disfluencies or repairs. These phenomena occur with great frequency in normal, spontaneous speech (e.g. [1, 2]), and yet we as listeners rarely seem to notice them. Researchers who have to transcribe normal speech often report finding it hard to detect disfluencies, to transcribe them correctly and to place them correctly even when they are specifically listening for them or doing verbatim transcriptions. Computational models of speech understanding, on the other hand, will attempt to assign lexical descriptions to everything in the speech signal and then try to resolve anomalies via mainly syntactic information (e.g. [3, 4]). It seems that the human listener may have a very useful ability to avoid some of the processing problems suggested by computational models: many disfluencies may be "filtered" out before lexical or syntactic processing commences.

Previous work on the perception of speech with disfluency has shown that under certain conditions disfluency can be detected very early in the speech signal, even before the recognition of the first word after the interruption, and that prosodic information might play an important part [1, 5, 6, 7]. But under more normal listening conditions, it may be that many disfluencies are missed altogether. Some previous work has discussed the phenomenon we examine here to a limited extent. Martin and Strange (1968) [8] suggested that listeners noticed very few disfluencies and tended to displace the few that they heard towards clause boundaries. Duez (1993) [9] found that "prepausal" lengthening was a valuable cue to the *detection* of self-interruptions in French speech. But in neither case was much distinction made between types of disfluency. The eventual aim of the present study is to discover, via perceptual experiments, which types of disfluency are most likely to be missed by listeners and to look for explanations for this perceptual illusion. The work described in this paper is a first step towards this aim. We also describe a larger project for which this work can be seen as a pilot study

CORPUS

The spontaneous speech materials used for the experiments were a set of 6 instructional monologues in Dutch collected and transcribed orthographically by Blaauw [10] and based on an idea by Terken [11]. In each monologue, a different male native speaker of Dutch described the construction of a picture of a house from pieces of coloured card to a listener who was neither visible nor audible to him. The length of the six monologues varied from about 3 minutes to about 17 minutes, depending on the amount of detail each speaker considered necessary for successful completion of the task.

Disfluencies were marked in the transcription by the author and checked by two other phoneticians.

The corpus consisted of a total of 4885 words in which 762 disfluencies, including silent pauses, filled pauses, repetitions and false starts, were identified. Disfluencies thus occurred every 6.4 words overall, with rates for individual speakers ranging from every 5.3 words to every 9.6 words. The disfluency types which concern us for the present study are filled pauses (N = 208), repetitions (N = 65) and false starts (N = 96).

EXPERIMENT

Materials

The speech materials used for the experiment were the full recordings of the six monologues described above. These were presented over stereo headphones from DAT tapes. The original transcriptions of the monologues were edited so that all disfluencies were removed as well as all labels and then printed out with double spacing between lines and triple spacing at major instruction boundaries to make it easier for subjects to follow them. In addition to the transcriptions, several pieces of coloured card were provided, which could be used to form a picture of a house.

Subjects and Procedure

Twenty subjects were paid to take part in the experiment. All were native speakers of Dutch, students or staff of the University of Utrecht, between the ages of 18 and 30, who reported no hearing defects.

Subjects were seated in a soundproof booth and given an instruction sheet to read, describing the task. When it was clear that the subjects understood the instructions clearly, the experiment began. Subjects were asked to listen to the tape and follow the transcriptions, marking with a cross any point at which the speech and the script differed. At the same time, subjects carried out the house-building task described in the monologues: this ensured that they were actually attending to the meaning of the speech, rather than concentrating fully on spotting anomalies, and thus made the listening task more realistic.

At the end of each monologue, the tape was stopped and the house that the subject had built was examined. Each subject was tested individually. The running time for the experiment was about 50 minutes.

Results

For each filled pause, repetition and false start, the number of subjects detecting the disfluency was totaled and these totals were then averaged for each type of disfluency. In many cases the outcomes for an individual disfluency were confused by the adjacency of another or by the fact

that one occurred within another (a filled pause or a repetition might be found within a false start, for example) so it was impossible to decide which to count as having been detected or missed. For this reason, from the original totals of disfluencies found in the corpus, we were left with 125 "clean" filled pauses, 16 "clean" single word repetitions, insufficient numbers of other repetitions, and 29 "clean" false starts.

Subjects were able to detect filled pauses 55.2% of the time. A clear difference was found between detection of filled pauses which were *within* sentences and those *between* sentences. Within sentences, filled pauses were correctly spotted 51.4% of the time (N = 88), while between sentences they were more easily detected (65.4%, N = 37).

A difference in the detectability of repetitions and false starts is suggested by the outcomes for those with single-word reparanda: single-word repetitions were detected only 27% of the time (N = 16), where false starts of the same length (in number of words) were detected at a rate of 39.3% (N = 7). In addition, longer false starts appeared to be easier to spot, with a 50% success rate (N = 22) for those with a reparandum of two words or more. Note that longer and more complex disfluencies were excluded from the analysis because of their complexity but that their rate of detection could be estimated at 90-100%.

DISCUSSION

A transcription-checking experiment tested the ability of listeners to detect disfluencies in spontaneous speech.

All types of disfluency included in the analysis showed lower than optimal rates of detection, ranging from 27% for single-word repetitions to 65.4% for filled pauses.

Differences were found between certain types. Filled pauses between sentences were easier to detect than those within sentences. There are a number of possible explanations for the difference, which will be investigated further in later work: filled pauses between sentences may simply be acoustically more prominent than those within either because of features of the filled pauses themselves (e.g. mid-clause filled pauses have pitch features which vary with their context [12]) or because of a more prominent pause context [9]; cognitive processing load on the listener may be lighter between sentences, allowing greater attention to the detection task, rather than to understanding the message. Single-word repetitions were harder to detect than false starts of the same length: one possible explanation for this is that "clean" repetitions are likely to be "prospective" and thus less likely than others to be accompanied by pause [13] which has been suggested as a possible cue to detection [9]; another explanation may be that since such repetitions are most likely to be short function words [1], they will be less prominent acoustically and perceptually than the words occurring in false starts of the same or greater length. Finally, the expected result that disfluencies with longer reparanda would be easier to detect was confirmed.

The number of "clean" tokens available in this corpus makes it difficult to come to firm conclusions on these results alone. However, the indications suggested here provide useful seeds for further study, which will involve a considerably larger amount of data.

WORK IN PROGRESS

The work described in this paper provides useful input to a larger project currently underway at the University of Edinburgh. Using the HCRC map task corpus [14] and performing a series of transcription experiments and acoustic and prosodic analyses we investigate the relationship between the tendency of disfluencies *not* to disrupt the processing of speech and the acoustic and prosodic features of such speech. It is hoped that this research will be of great value to our general understanding of how listeners process spontaneous speech.

ACKNOWLEDGEMENT

This work was carried out while the author was working at the Research Centre for Language and Speech (OTS), Utrecht University, The Netherlands. The author wishes to thank Noortje Blaauw, Sieb Nooteboom and Hugo Quené for their help and advice on the project. The current project at the University of Edinburgh is supported by ESRC Award No. R0002352.

REFERENCES

[1] Lickley, R.J. (1994), *Detecting disfluencies in spontaneous speech*, Unpublished PhD thesis, University of Edinburgh.

[2] Shriberg, E.E. (1994), *Preliminaries to a Theory of Speech Disfluencies*, UC Berkeley.

[3] Hindle, D. (1983), *Deterministic parsing of syntactic non-fluencies*, in Proceedings of the 21st annual meeting of the Association for Computational Linguistics, pp 123-128.

[4] Shriberg, E.E., Bear, J. and Dowding, J. (1992), *Automatic Detection and Correction of Repairs in Human-Computer Dialog*, in Proceedings of the DARPA Speech and Natural Language Workshop. Marcus, M. (ed.).

[5] Lickley, R.J., Bard, E.G. and Shillcock, R.C. (1991), *Understanding Disfluent Speech: is there an Editing Signal?*, Proceedings of the ICPhS, Aix-en-Provence, France, vol 4 pp 98-101.

[6] Lickley, R.J. Shillcock, R.C. and Bard, E.G. (1991), *Processing Disfluent Speech: How and When are Disfluencies Found?*, Proceedings of Eurospeech 91, Genova, Italy, pp 1499-1502.

[7] Lickley, R.J. and Bard, E.G. (1992), *Processing Disfluent Speech: Recognising Disfluency Before Lexical Access*, Proceedings of The ICSLP, Banff, Alberta, Canada, pp 935-938.

[8] Martin, J.G. and Strange, W. (1968), *The Perception of Hesitation in Spontaneous Speech*, in Perception and Psychophysics 3(6), pp 427-438.

[9] Duez, D. (1993), *Acoustic correlates of subjective pauses*, in The Journal of Psycholinguistic Research 22(1), pp 21-39.

[10] Blaauw, E. (1994), *The contribution of boundary markers to the perceptual difference between read and spontaneous speech*, in Speech Communication 14, pp 359-375.

[11] Terken, J.M.B. (1984), *The distribution of pitch accents in instructions as a function of discourse structure*, Language and Speech 27(3), pp 269-290.

[12] Shriberg, E.E. and Lickley, R.J. (1993), *Intonation of clause-internal filled pauses*, Phonetica 50, pp 172-179.

[13] Hieke, A. (1981), *A content-processing view of hesitation phenomena*, Language and Speech 24(2), pp 147-160.

[14] Anderson, A.H. *et al.* (1991), *The HCRC map task corpus*, Language and Speech 4, pp 351-366.

SENTENCE COMPREHENSION AND TEXT COMPREHENSION: CHILDREN'S STRATEGIES

M. Gósy
Research Institute for Linguistics, Budapest, Hungary

ABSTRACT

The aim of this paper was to investigate the strategies children use when comprehending sentences vs. texts. 100 preschool girls and boys took part in the experiments. The results reveal a number of strategies the children use when showing age-required performance or when underperforming. Children of this age have better developed strategies for sentences than for texts, and their operations at the highest level seem to be restricted.

INTRODUCTION

The various levels of the speech decoding process are differently involved in a comprehension task and are assuemed to be activated according to the actual speech in-put. Participition of the higher levels (comprehension, associations) depends on the complexitiy and size of the speech input, i.e. on the semantic and syntactic contents [1]. After the successful analysis of a sentence without any context, comprehension might take place without the activation of the level of associations. However, full comprehension of a text usually involves the operations of the highest level as well. This difference explains the cases of good understanding of sentences when having problems with text comprehension, and the cases of good comprehension of texts when having problems with sentence understanding. The latter appears e.g. in the key-word-strategy phase of first language acquisition when the child's decoding mechanism operates with familiar words to understand longer text-like utterances. This strategy works with the activation of associations where these operations "replace" the actual lexical and syntactic access the child would have needed [2]. On the contrary, despite acceptable sentence comprehension children sometimes are not able to figure out the text cohesion, to realize the semantic interrelations within sentences and the semantics of these interrelations, i.e. to comprehend the text. From the aspect of speech comprehension, it is obvious that this process can work with or without problems since the output is defined by the necessary levels involved in the operations.

Questions have arisen concerning the comprehension strategies of preschool children since great differences had been found in their performances. A series of experiments has been carried out to answer the following questions: (i) What are the strategies Hungarian preschool children use when understanding sentences and texts? (ii) What are their problems in the comprehension tasks? (iii) How do our results relate to the language acquisition process of Hungarian 6-year-olds?

METHOD AND MATERIAL

For the sake of this experiment an immediate off-line method of the GMP standardized Hungarian test [3] has been chosen. Sentence comprehension was checked by using colour pictures. 10 sentences with various semantic and syntactic structures were created focusing on four critera. (i) Those word classes, morphological and syntactic structures were selected that appear latest in the Hungarian-speaking children's speech production. (ii) Those semantic and lexical units were preferred that occur in children's speech production at the examined age relatively rarely. (iii) All sentences should be stored and reproduced easily. (iv) All sentences and the opposite of their semantic content were to be easily represented in a picture. The semantic difference of these pairs of sentences was to be demonstrated by one visual difference in the picture. The size of the sentences were similar taking into consideration the operation of the short-term memory. E.g.: *The girl must give the book to the boy.* (In the picture the boy gives the book to the girl.).

After showing the two pictures (one for the target and another for the opposite sentence) the examiner uttered the target sentence to the child whose task was to choose one of the two pictures appropriate to the utterance heard.

For the text comprehension task, a short story about animals was used that had been recorded by a male voice. The total duration of the story was 1.15 minutes. The speech tempo of the speaker was 10.2 sounds/s on average (i.e. slower than the adults' average). Ten comprehension questions were created concerning the details and the interrelations of the text (wh-questions). The child's task was to listen to the story and to answer the examiner's 10 questions.

100 children were tested individually from three ordinary Hungarian kindergartens with heterogeneous social background. Those 6-year-olds were selected for the experiment who were going to start the school the next school-year. Their ages were between 6;0 and 6;11: 51 girls and 49 boys.

RESULTS

Results show that children's sentence comprehension is better than their text comprehension in all subgroups of the tested 6-year-olds (Fig. 1). This means that the activation of the level of associations seems to be difficult for the majority of children, however, there are differences in the performances across subgroups in both tests (Tables 1 and 2).

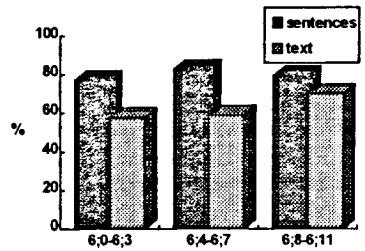


Figure 1. Average values of sentence and text comprehension in the age-subgroups.

Table 1. Data of correct answers of six-year-olds in sentence understanding test.

Ages	Correct sentence comprehension average (%) girls/boys	range (%) girls/boys
6;0-6:3	82.38/71.66	50-100/40-100
6:4-6;7	85/80.52	50-100/50-100
6;8-6;11	78/80.71	50-90/50-100
Average	81.79/77.63	(79.71)

Comparisons have been made to the standard value that is minimum 80% correct comprehension of all sentences for this age. According to the average data our preschool children show some backwardness; particularly the youngest boys and the oldest girls. The range of children's correct performance is relatively wide. 70.62% of all children have reached the expected level (their results show 80% or better correct performance in this test). However, almost 30% of all children have performed extremely poorly. Their performance is equal to the sentence comprehension of the normally developed Hungarian-speaking 5-year-olds.

Text comprehension was much worse with the tested children; according to the standardized data (GMP). A significant difference ($p<0.001$) has been found

between the standard value of 6-year-olds' text comprehension (minimum 70% correct answers) and our 6-year-olds' performance (average: 62.3%). For the speech perception process it is supposed that almost 50% of our tested children either (i) are unable to activate the level of associations during comprehension or, (ii) their operations at this level are false and/or ambiguous. 29.92% of all children could answer 5 or less questions correctly while 16.72% of all children performed according to the standard performance of the normally developed Hungarian-speaking 3- and 4-year-olds.

Table 2. Data of correct answers of 6-year-olds in text-comprehension test.

Ages	Correct text comprehension	
	average (%) girls/boys	range (%) girls/boys
6;0-6;3	60/55.5	20-90/10-90
6;4-6;7	52.77/65.2	10-90/30-100
6;8-6;11	71/69.28	40-90/40-100
Average	61.25/63.36 (62.3)	

There are only three sentences where the false answers were relatively frequent: one concerning a semantic unit, another one concerning morphological homonyms with diverse semantics and the third one concerning a syntactic structure. There have been false answers in 25% of all responses for the sentence: *The mouse has almost reached the cheese*. Here, the ambiguous interpretation of the word meaning 'almost' is the reason for the false responses. For the second case 44% of all answers were incorrect which had been caused by the homonymous morphological structure of the dative suffix *-nak* 'to' and that marking the subject of the verb 'must' (cf. the sentence in English: *The girl must give the book to the boy.*). For the third sentence, *Before drinking, the bear had eaten something.*, 61% of all responses were incorrect.

Analysis of children's text comprehension shows that there were only 4 questions answered correctly in 70% or more. All of them concerned the details of the heard text. 50% or more of all answers were false for the two following questions. Both of them concerned details presented at the very beginning of the story. The answers for these questions with the majority of children show the unnecessary activation of the level of associations. Instead of the right answers the children tried to give a structurally adequate but semantically inappropriate answer.

INTERRELATIONS OF SENTENCE COMPREHENSION AND TEXT COMPREHENSION

There can be a very clear explanation for the equal performance of children in both tests independently of the correctness of their interpretation. The ambiguous and/or false or, the unambiguous and good operations at various levels of the decoding process lead to an equality of performance: the child either comprehends speech (both sentences and texts) without any problem in semantics and in syntactic relations or, the child fails because of distorted working of his processes. The interesting quesions in this latter case are: which are these operations and how much are they distorted?

There were children who showed different performance depending on the speech input. Two different types have been found: (i) the child's performance meets the age requirements in sentence comprehension but not in text comprehension, and (ii) the child's performance meets the age requirements in text comprehension but not in sentence comprehension.

12 out of 100 children's were found to perform better in text comprehension than in sentence comprehension. These children are able to use and activate the necessary associations immediately before completing the lexical access since their operations here are ambiguous. This performance can be understood as a transformation of the former 'key-word-strategy' at a higher cognitive level. The essential difference is that the original key-word-strategy functioned within an age-characteristic decoding mechanism where total lack of certain semantic and syntactic knowledge was substituted by the comprehension of interrelations of key words. However, at the age of 6 there is an ambiguous knowledge concerning certain semantic and syntactic units and this uncertainty is substituted by an attempt to comprehend the interrelations of supposedly comprehended items within a text. Those children, whose sentence comprehension was good, succeeded (this is our 12 children) but those children whose sentence comprehension was below the age-requirements were unable to use this strategy (17 children).

31 children performed well in the sentence comprehension task while they underperformed in the text comprehension task. What strategy could these children use? It can be claimed that the speech decoding process of these children operates well up to the level of associations. It means that they are able to be successful with lexical access, and they are able to identify both the semantic and syntactic interrelations across a small number of words. However, they are unable to find the connections among certain items of a longer speech input, a text. For the reason of that two possibilities suggest themselves. (i) Their speech perception processes work slowly because each level needs too much information to operate and this time consuming working does not allow the mechanism to operate at the highest level as well. (ii) The level of associations with these children may be unmatured, i.e. the operations are similar to those of children of younger ages. It is likely that both explanations are correct in the sense that there are children of whom (i) and others of whom (ii) is characteristic.

DISCUSSION

1. Our results have confirmed that (i) there is a significant difference between sentence and text comp-rehension of the tested 6-year-olds ($p<0.001$), and (ii) that sentence comp-rehension is better than text comp-rehension. Children's decoding strategies at the tested age are better for sentences than for texts.

2. No significant difference has been found either among the age-subgroups' performances or between girls and boys. There is one exception: there are 11 boys out of those 17 children who underperformed in boths tests.

3. Comparing our data to the standard values shows that 17% of all children are risk children for learning to read and write [4].

4. Various strategies have been found for solving the sentence and text comprehension tasks involving partly or completely the necessary levels of the decoding process. The strategy the child is supposed to use has also been supported by the correctness difference in the two types of tests used.

REFERENCES

[1] Altmann, G.T.M., ed. (1990), *Cognitive Models of Speech Processing*, Cambridge, Mass.: The MIT Press.

[2] Gósy, M. (1992), *Speech Perception*, Frankfurt: Hector.

[3] Gósy, M. (1995), *GMP-diagnostics* /in Hungarian/, Budapest: Nikol.

[4] Perfetti, C.A. (1987), "Language, Speech, and Print: Some Asymmetries in the Acquisition of Literacy", In: Horowitz, R. and Sammels, S.J. (eds.) *Comprehending Oral and Written Languages*. New York: Academic Press, pp. 355-370.

A DATA BASE TO EXTRACT PROSODIC KNOWLEDGE WITH A NUMERICAL AND SYMBOLICAL LEARNING SYSTEM

J.J. Schneider and H. Méloni
Laboratoire d'Informatique d'Avignon, France

ABSTRACT

In this paper we present a representation language to conceive prosodic data bases to extract prosodic knowledge in explicit rules form with a learning system working with numerical and symbolical information.

The multiplicity of parameters at work (acoustic, linguistic, syntactic, semantic, pragmatic, ...) make the realisation of a prosodic labelling complex. In the first time, we present the common basis principles of the prosodic models, and we define a representation language suited to this models. In the second part of this paper we present, in succinct way, the general methodology and tools to extract the prosodic knowledge. And, to illustrate our aims, we propose a prosodic labelling of the utterance to realise a prosodic data base on which we use the previous tools.

PROSODIC LABELLING - METHODOLOGY

All prosodic models, in sprite of their differences, have two common basis principles : (a) the melodic, rhythmic and intensity continuum could be segmented into discreet units, (b) the great congruence between the organisation of the utterance and prosodic organisation of the message [1][2][4].

THE INSTANCE LANGUAGE

For each unit, we combine a description of the prosodic parameters with a description of the organisation of the utterance. To realise this language we define the following descriptions.

An Elementary Description

The smaller description unit, called elementary description, is a pair (attribute value). Attribute is a symbolic label whose meaning is known by the expert, and value is an homogeneous data list which characterise the elementary description.

Example 1 : elementary descriptions

```
(PointA (+1.1 -3.05 +5.2))
(ValeurMax (128))
(Couleur ("bleu foncé"))
```

A simple data list don't exactly characterise an elementary description. For each one [3], we combine a type, a nature and an environment defined as following.

The type is the set of possible values. We define two types, the first one is numerical to describe crude data, and the second is symbolical to describe abstract concepts.

To describe the data behaviours we define three natures. The atomic descriptor's values are independent each other, the linear descriptor's values belong to a completely arranged set, and the structured descriptor's values belong to a generalisation tree.

The environment contain several information, one to describe the data presentation after organisation, and one to describe the list of mathematics tools needed.

The Description of a Parameter

For a best organisation, we define a parameter to group some elementary descriptions together.

```
(Energie
  (AlterationE (+1.3))
  (ContourE (71 66 65)))
(FrequenceFondamentale
  (ContourLocalF0 (167 120 137))
  (ContourGlobalF0 (135 120 116))))
```

The Description of an Instance

Two elements compose an instance, the description of the prosodic parameters, and the description of the organisation of the utterance. Each element of an instance is a list of descriptions of parameters.

```
( ( (FrequenceFondamentale
       (ContourLocalF0 (167 120 137))
       (ContourGlobalF0 (135 120 116)))
    (Duree
       (AlterationD (-11.4)))          )
  ( (StructureGrammaticale
       (Mot ("verbe"))
       (Groupe ("groupe verbal")))     ) )
```

The Data Base

An instances succession compose the data base, in which all instances must have the same structure.

DATA ORGANISATION - KNOWLEDGE ACQUISITION

In the first time, we regroup all instances in which the organisation of the utterance is exactly the same. This step could be supervised by an expert according to the clustering technical used. We obtain a representation more concise and structured of the data. To express this result we propose a set of possible representations.

The next step determine, with a generalisation on the set of the descriptions of the organisation utterance, a characterisation of the phenomena observed on the corpus used.

PROSODIC LABELLING

The Corpus

The corpus chosen to validate our approach had been conceived by V. Aubergé and G. Bailly within the context of the MULTIDIF 1991 contract.

The Description Unit

In this paper the description unit chosen is the vowel. It could be easy extended to any other unit, like syllable or word, according to the phenomena looked for.

The Prosodic Parameters

We propose for the prosodic parameters (fundamental frequency, duration and intensity) a set of marks and a representation of them.

Fundamental Frequency

To underscore the local and global phenomena we define two elementary descriptions, the first one called local fundamental frequency, and the second one called global fundamental frequency. For each description we propose three points defined in the *Figure 1* and *Figure 2*.

Figure 1: local fundamental frequency

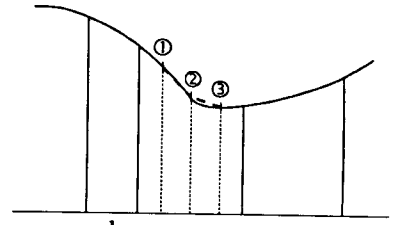
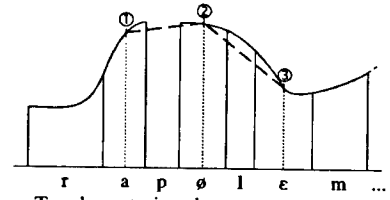


Figure 2: global fundamental frequency



To characterise these movements we use the variation rate between the three points, in a height values scale [4]. The following example is a description of the prosodic parameter fundamental frequency.

```
(FrequenceFondamentale
  (ContourLocalF0 (-3 -1))
  (ContourGlobalF0 (+1 -3)))
```

Duration

For each vowel we could establish the intrinsic duration [5]. We calculate with (1) an alteration rate between the measured and the theoretical duration.

$$\%D(V) = \frac{d(V)}{d_{th}(V)} - 1 \qquad (1)$$

with %D(V) duration alteration rate
d(V) measured duration
$d_{th}(V)$ theoretical duration

We don't calculate the intrinsic duration of each vowel but we calculate an average duration. If vowels are in varied linguistic contexts and if the corpus is big enough, this value is satisfactory.

$$d_{th}(V) = \frac{1}{N}\sum_{i=1}^{N} d_i(V)$$

with $d_{th}(V)$ theoretical duration
N total number of apparitions
$d_i(V)$ i^{th} measured duration

To express this alteration rate we use a symbolical elementary description defined in the *Table 1*.

Table 1 : relations between numerical and symbolical values.

[... -10%]	[-10% 10%]	[10% ...]
"réduite"	"normale"	"allongée"

The following example is a description of the prosodic parameter duration.

```
(Duree
  (AlterationD ("réduite")))
```

Intensity

To describe the prosodic parameter intensity we use two elementary descriptions. The first one, like duration, is an alteration rate between the theoretical intensity and the measured intensity of the vowel.

$$\%E(V) = \frac{e(V)}{e_{th}(V)} - 1 \qquad e_{th}(V) = \frac{1}{N}\sum_{i=1}^{N} e_i(V)$$

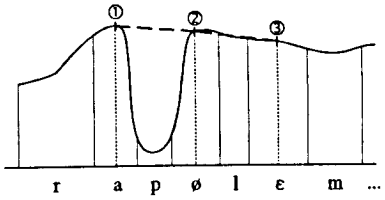
To express this alteration rate we use a symbolical elementary description defined in *Table 2*.

Table 2 : relations between numerical and symbolical values.

[... -5%]	[-5% 5%]	[5% ...]
"basse"	"normale"	"élevée"

The second elementary description, like fundamental frequency, underscore the relative intensity evolution between the previous vowel, the current vowel and the follow vowel in the sentence. We propose three points defined in the *Figure 3*.

Figure 3: intensity movement



To characterise these movements we use the variation rate between the three points, in a height values scale. The following example is a description of the prosodic parameter intensity.

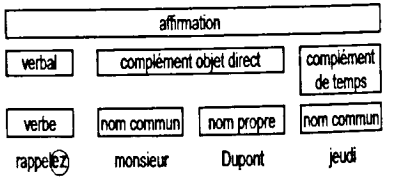
```
(Energie
  (ContourE (-1 -2))
  (AlterationE ("normale")))
```

The Utterance Organisation Parameters

The Syntax

In French the syntax is the most influent utterance organisation parameter.

Figure 4 : syntactic tree representation



From the syntactic tree (see *Figure 4*), we define two parameters. The first one describe the grammar nature of each node of this tree. The second describe the position in the tree of each constituent in the upper constituent. The following example is a description of the grammar structure and the positions structure.

```
(StructureGrammaticale
  (Mot ("verbe"))
  (Groupe ("groupe verbal")))
(StructurePositions
  (Voyelle (3 3))
  (Mot (1 1))
  (Groupe (1 3)))
```

The Phonetic Context

To underscore the influence of the phonetic context on the prosodic parameters we propose three parameters. Two for the phonetic description of the adjacent consonants and one for the phonetic description of the current vowel. For each parameter we use two elementary descriptions, one for the nasality and one for the opening (for a vowel), and one for the voicing and one for the articulatory mode (for a consonant). The following example is a description of the phonetic context.

```
(ConsonnePrec
  (ModeCP ("vocalique"))
  (VoisementCP ("voisée")))
(ConsonneSuiv
  (ModeCS ("occlusive"))
  (VoisementCS ("non voisée")))
(Voyelle
  (NasaliteV ("orale"))
  (OuvertureV ("ouverte")))
```

The Pause

To underscore the influence of the pause on the prosodic parameters we propose two parameters. The first one describe the distance to the next pause, expressed in milliseconds, and the second describe the duration of this pause expressed milliseconds to. The following example is a description of the pause.

```
(Pause
  (DistancePause (430))
  (DureePause (90)))
```

RESULTS - CONCLUSION

We have made tests on a corpus labelled in a voluntary simple way and the results obtained are very interesting.

With this language we can describe the general phenomena found in the prosodic models. The prosodic data base presented could be extended to any other parameters like semantic, pragmatic, ...

The originality of our approach is the system only know the structure of an instance. The segmentation unit and the description of the prosodic parameters and the organisation of the utterance parameters could be different according to the corpus and the phenomena looked for. In the same way, clustering and generalisation tools are used like resources, and they could be different.

In conclusion, the most interesting aspect of the environment presented is his opening, to the phenomena descriptions, to the description unit and to the techniques.

REFERENCES

[1] J.J. Schneider P. Robineau and H. Méloni "*Un système d'Apprentissage Symbolique et Numérique pour la Formalisation de connaissances Prosodiques*" XX^{ème} Journées d'Etude sur la Parole - Trégastel 1994

[2] L. Mortamer F. Emerard and L. Miclet "*Attempting automatic prosodic knowledge acquisition using a database*" Workshop on Speech Synthesis - Autrans 1989

[3] R.S. Michalski "*A Theory and Methodology of Induction Learning*" Machine Learning : An Artificial Intelligence Approach vol. 1 - 1983

[4] G. Caelen-Haumont "*Stratégie des locuteurs en réponse à des consignes de lecture d'un texte : analyse des interactions pragmatiques et des paramètres prosodiques*" Thèse de Doctorat d'état 1991

[5] Y. Nishinuma S. Barber and D. Hirst "*Estimation de la durée intrinsèque des voyelles*" XII^{ème} Journées d'Etude sur la Parole - Montréal 1981

THE LABELLING OF PROMINENCE IN SWEDISH BY PHONETICALLY EXPERIENCED TRANSCRIBERS

Eva Strangert and Mattias Heldner
Department of Phonetics, Umeå University, Sweden

ABSTRACT

An IPA-based system has been agreed upon for labelling Swedish prosody. In the present study this system is evaluated by assessing the inter-transcriber reliability in prominence labelling of nine expert subjects. The study also explores the acoustic (F0) basis for observed variability in the assignment of focus accent, the highest prominence label.

INTRODUCTION

Recently, as large corpora of prosodically labelled speech are needed for quantitative computational modelling of speech, great efforts are being taken to develop transcription systems meeting high standards on reliability. Thus, before extensive use of a system is initiated, it must be evaluated. The TOBI (TOnes and Break Indices) system developed for transcribing English prosody has been evaluated in a number of studies eg. [1,2]. Reyelt [3] evaluated a number of variants of prosodic transcription for German within the VERBMOBIL project. For Swedish, an IPA-based system has been agreed upon for labelling prosody (prominence and boundary phenomena), the details of which have been described in [4]. We have used this system in two studies [5,6] comparing the labelling of boundaries and prominences in spoken Swedish made by phonetically experienced and non-experienced transcribers.

In the present study, the scope has been widened. One purpose, which it shares with the former studies [5,6], is to evaluate the transcription system used for labelling. In particular we want to estimate the extent to which experienced phoneticians and speech researchers vary in their labelling of prominences when presented with samples of read and spontaneous Swedish. In addition, the study aims at exploring the acoustic basis, specifically F0-characteristics, for the variability in labelling that we predict will occur. In particular, we want to establish the extent to which the variability associated with the assignment of focus accent is explainable in terms of F0-cues.

Beckman [7] reviews the research on acoustic correlates to perceived stress in English. Referring to study [8], Beckman [7, p 60-62] makes clear that the dependence of perceived stress on F0-cues is complex, and varies with the position of the word in the sentence. Further, Wells [9] concludes that F0-cues play an important role for perceived prominence in English, although various other cues contribute, too. Although F0 is not assumed to be the only cue to prominence in Swedish – Bruce [10] also mentions temporal correlates, and there are also data reported in [11] indicating temporal correlates – it is believed to be an important determiner of focus accent. Thus, relating perceived focus accent to F0-events seems reasonable in the light of previous research [12] according to which focus accent is intimately tied to a F0-rise following a word accent F0-fall timed differently for words with acute and grave accent, respectively.

EVALUATION OF THE TRANSCRIPTION SYSTEM

Method

The 9 subjects participating in the study are all phoneticians or speech researchers with wide experience in prosody from different sites in Sweden. All are native-born Swedes.

The subjects transcribed two kinds of recorded speech material. One was an excerpt, 233 words long, from an authentic news cable read aloud. The other was a 252-word-long excerpt of spontaneous speech, a retelling of the story read aloud. Both recordings were made in a soundproof room and rendered by the same male Swedish speaker.

Each expert was sent the recorded material and instructions for labelling prominence according to the IPA-based Swedish system. Following this, four levels of prominence were distinguished and labelled accordingly for each word in the material: no stress (unmarked), secondary stress (,), primary stress/accented (') and focus acccent (").

Subsequent analyses included coding the data (no stress=0; secondary stress=1; primary stress=2; focus accent=3) and statistical analyses to estimate reliability.

Labelling data

Table 1 shows the labelling of prominences by the nine experts in a sample of the read material. The words in the text are ordered vertically in the first column. The following nine columns contain the individual labellings of the transcribers and the tenth column the means of these labellings for each word. The data presented give a rough indication of the reliability of labelling.

Table 1. Labelling by nine transcribers. 0=no stress, 1=secondary stress, 2=primary stress, 3=focus accent.

Word	Transcribers 1 — 9									$\bar{X}$
enligt	0	0	1	0	0	0	0	0	0	0.1
libyska	3	2	3	3	3	3	2	3	3	2.8
uppgi...	2	2	2	2	2	2	2	2	2	2
föll	0	0	2	0	0	1	0	0	0	0.3
åtta	2	3	3	2	3	3	2	3	2	2.6
450-k...	2	2	3	3	3	2	2	2	2	2.3
över	0	0	1	0	0	0	0	0	0	0.1
Tripoli	2	2	3	3	3	3	2	3	3	2.7
och	0	0	0	0	0	0	0	0	0	0
Bengazi	2	3	3	2	3	2	2	3	2	2.4
när	0	0	0	0	0	0	0	0	0	0
de	0	0	0	0	0	0	0	0	0	0

Inter-transcriber reliability

Generally, reliability concerns the extent to which measurements are repeatable in a variety of conditions. Within this framework we will consider two aspects, the one concerning the extent to which the transcribers covary, that is, give relative labelling values that are correlated, and the other concerning the extent to which the transcribers give identical labels. We will henceforth refer to the first as 'reliability' and the second as 'agreement'. All computations are made with acute and grave accent words pooled.

The inter-transcriber reliability (Cronbach's alpha) for prominence is .98 for read and .97 for spontaneous speech (difference not significant). That is, the transcriptions are highly reliable in the sense of relative labelling consistency irrespective of the material.

To determine the reliability in the more strict sense of agreement, that is identical matching, we used the same test as Silverman et al. [1] and Pitrelli et al. [2]. They calculated the agreement across all possible pairs of transcribers for each word of each utterance labelled. The index was calculated as the average percentage of agreeing pairs and, according to the criterion set in [1], the agreement should be at least 80%. Calculated on our data, this index is 78% and 71%, for the read and spontaneous speech respectively, thus indicating a somewhat higher agreement on the read speech. There are several differences between TOBI and our system which make comparisons complicated. For the TOBI transcribers, the task was to decide whether a word had a pitch accent or not, and if so, what kind of pitch accent. The indices reported for these tasks were 86% and 64% respectively for the 4 most experienced of their 20 transcribers.

We also calculated an index estimating the extent to which *all* the transcribers made *exactly* the same judgements on each word. A detailed account of these calculations and other evaluation data presented here are given in [6].

F0 IN RELATION TO PROMINENCE LABELS

Method

The subsequent analysis was made on 60 acute and 55 grave accent words judged to be focussed (that is, having a prominence degree of 3, according to our coding) by two or more of the nine transcribers. For each of these words a prominence mean score based on the labelling of all nine transcribers was calculated.The words were digitized at 44.1 kHz. Measurements were made in both the read and spontaneous speech of the size of the word accent fall and the focus accent rise.

To calculate the falls and rises four measuring points were defined, primarily on the basis of the F0 tracings, see the illustrations in Figure 1: (1) The beginning of the word accent fall; the highest point in the word accent fall. (2) The end of the word accent fall; the lowest point in the word accent fall. (3) The beginning of the focus accent rise; the lowest point in the focus accent rise. For acute accent words this point coincides with (2). For grave accent words it either coincides with (2) or, in the case of longer words, may be located at som distance from (2).

(4) The end of the focus accent rise; the highest point in the focus accent rise. In a few cases in which the critical F0-events were not easily located, additional criteria were used, determined on the basis of the patterns observed in the unequivocal cases. We also used [13] as a reference when deciding on these additional criteria.

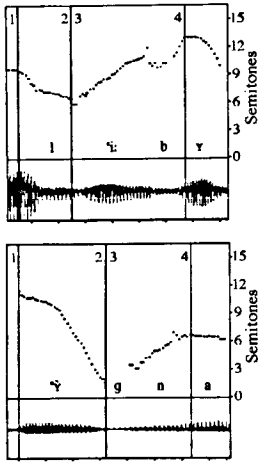


Figure 1. Measurement points. Above : underlined portion of libyska (acute accent); below: underlined portion of byggnad (grave accent).

The word accent fall is defined as the difference between points (1) and (2) and the focus accent rise is defined as the difference between points (3) and (4) measured in semitones. In addition we tested two other F0-parameters, differences (focus accent rise–word accent fall) and ratios (focus accent rise/word accent fall).

Results

The majority of the prominence mean scores for all acute and grave accent words included in the analysis fell in the range between 2 and 3. (It should be recalled that a word judged to be focussed is coded as 3 in our analysis. Therefore, mean scores close to 3 indicate a general agreement on the word as being focussed.) The prominence mean scores were then used in multiple regression analyses to determine if, and to what extent, the measured F0 movements (with word accent fall and focus accent rise as the independent variables) could explain the variability in the prominence scores.

The results demonstrate insignificant effects of the word accent fall in the read as well as the spontaneous speech and for words with acute and grave accent alike. The focal accent rise, on the other hand, is significantly correlated with perceived scores ($p<.05$) both in the read

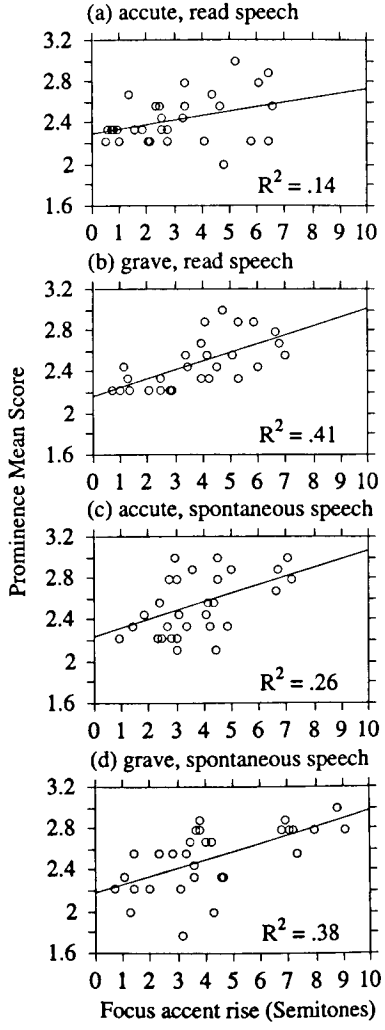


Figure 2. Regression analyses of size of focus accent rise and prominence mean score for 60 acute and 55 grave accented words in read and spontaneous speech.

and spontaneous speech and for acute and grave accent words (Figure 2 a-d). That is, the greater the size of the rise, the stronger the agreement on focus accent. Both kind of data therefore corroborate previous results demonstrating greater effects on perceived prominence of the rise than the fall [11,12]. However, the R-square values, correlations in terms of explained variance, are quite low for all four regression models, .14 and .26 for the acute accent and .41 and .38 for the grave, indicating other influences than F0 on perceived prominence cf. [11].

We also did regression tests with differences as well as ratios between the focus accent rise and the word accent fall as independent variables, but neither of them reached significance.

CONCLUSIONS

In this prosodic transcription evaluation we have demonstrated the capacity of the system as used by expert transcribers. The reliability is high as well as the inter-transcriber agreement. Exploring the acoustic basis for observed variability associated with the assignment of focus accent, we found that the greater the F0-rise, the stronger the agreement on focus accent. That is, the size of the focus accent cues the degree of prominence. Yet it explains only part of the variation. In conclusion then, there are other important cues to perceived prominence (focus accent) than those investigated here. We are in the process of conducting a study including temporal as well as other cues to perceived focus accent.

ACKNOWLEDGEMENTS

We gratefully acknowledge the contributions of Gösta Bruce, Rolf Carlson, Claes-Christian Elert, Anders Eriksson, Gunnar Fant, Eva Gårding, Olle Kjellin, Anita Kruckenberg, Per Lindblad, Ulla Sundberg, without whose help this study would not have been possible.

This research was supported by grants from the Swedish HSFR/NUTEK Language Technology Programme.

REFERENCES

[1] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992), "TOBI: A standard for labeling English prosody". In *ICSLP 92 Proc.* pp. 867-870, Banff, Alberta, Canada.

[2] Pitrelli, J., Beckman, M. & Hirschberg, J. (1994), "Evaluation of prosodic transcription labeling reliability in the ToBI framework". In *ICSLP 94 Proc.* pp. 123-126. Yokohama, Japan.

[3] Reyelt, M. (1993), "Experimental investigation on the perceptual consistency and the automatic recognition of prosodic units in spoken German", Proc. ESCA Workshop on Prosody, *Working Papers*, 41, pp. 238-241, Dept. of Linguistics and Phonetics, Lund Univ.

[4] Bruce, G. (1994), "Prosodisk strukturering i dialog". In *Svenskans beskrivning* 20. pp. 9-23. Lund: Lund University Press.

[5] Strangert, E. & Heldner, M. (1994), "Prosodic labelling and acoustic data", *Working Papers*, 43, pp. 120-123, Dept. of Linguistics and Phonetics, Lund Univ.

[6] Strangert, E. & Heldner, M. (1995), "Labelling of boundaries and prominences by phonetically experienced and non-experienced transcribers", *PHONUM 3*, Dept. of Phonetics, Umeå Univ.

[7] Beckman, M.E. (1986), *Stress and Non-Stress Accent*, Dordrecht: Foris Publications.

[8] Nakatani, L. & Aston, C. (1978), "Acoustic and linguistic factors in stress perception", Unpubl. ms, Bell Lab.

[9] Wells, W.H.G. (1986), "An experimental approach to the interpretation of focus in spoken English", in C. Johns-Lewis (Ed.), *Intonation in Discourse*, 53-75, London: Croom Helm.

[10] Bruce, G. (1983), "Accentuation and timing in Swedish", *Folia Linguistica*, vol. 17, pp. 221-238.

[11] Sundberg, U. (1994), "Tonal and temporal aspects of child directed speech", Working Papers 43, pp. 128-131 Dept. of Linguistics and Phonetics, Lund Univ.

[12] Bruce, G. (1977), *Swedish Word Accents in Sentence Perspective*, Lund: CWK Gleerup.

[13] Engstrand, O. (1989), "Phonetic features of the acute and grave word accents: data from spontaneous speech", *PERILUS X*, pp. 13-37, Inst. of Linguistics, Univ. of Stockholm.

COMBINING STATISTICAL AND PHONETIC ANALYSES OF SPONTANEOUS DISCOURSE SEGMENTATION

Marc Swerts
Institute for Perception Research (IPO), Eindhoven, The Netherlands

ABSTRACT

This paper presents a method to study prosodic features of discourse structure in unrestricted spontaneous speech. Past work has indicated that one of the major difficulties that discourse prosody analysts have to overcome is finding an independent specification of hierarchical discourse structure, so that one avoids circularity. Previous studies have tried to solve this problem by constraining the discourse or by basing segmentations on a specific discourse theory. The current investigation explores the possibility of experimentally determining discourse boundaries in unrestricted speech. In a next stage, it is investigated to what extent boundaries obtained in this way correlate with specific prosodic variables.

INTRODUCTION

It is intuitively clear that most discourse exhibits structure in that it consists of larger-scale information units, or discourse segments. Those segments can be viewed as building a discourse hierarchy, since segments may be embedded in others: for instance, someone may talk about his holidays, with subtopics on hotel, food, and so on. In practice, however, it is often difficult to specify exactly the boundaries of those higher-level units and their mutual relationships. This poses a serious methodological problem to investigators who study linguistic, e.g., prosodic, correlates of discourse structure. Therefore, one would like to obtain ideally an 'independent' specification of information structure so as to avoid circularity. That is, it needs to be guaranteed that the junctures in the information flow are determined independently from prosodic considerations [1]. In the literature, one sees basically two solutions to overcome this problem.

In a first line of research, the problem is somewhat circumvented by looking at construed speech materials. One group of researchers has looked at read-aloud texts with predetermined paragraph boundaries (e.g., [9], [3], [6]); similarly, others have focused on tightly constrained types of spontaneous speech, by experimentally eliciting discourse in such a way that it becomes easily segmentable in consecutive information units ([8], [7]). In this way, prosodic features of discourse segments can be adequately investigated. These studies are limited, though, in that the structures investigated are controlled, but overly simple. It remains to be seen to what extent the findings can be generalized to more complex discourse.

The second approach is more theory-based in an attempt to motivate segmentations on the basis of explicit models of discourse structure. In studies such as [2] and [4], both within the Grosz and Sidner framework, 7 subjects were instructed to segment a set of monologues, using speaker intention as a criterion. It turns out that there is considerable variation between labelers as no two segmentations are the same. In particular the specification of hierarchical relationships between segments appears to be difficult. Therefore, it is decided in these studies either to concentrate on only those structural features agreed upon by all labelers [2], or retain those boundaries assigned by at least 4 out of 7 labelers [4].

This paper addresses another approach, partly inspired by Rotondo ([5]). Instead of taking the variance between labelers as a disadvantage, it rather exploits it to specify hierarchically different discourse boundaries. In contrast with earlier studies, it takes the segmentations of relatively many labelers to arrive at this goal. Basically, boundary strength is then computed as the proportion of subjects agreeing on a given break. In the following, it will be illustrated how this method offers a useful alternative to already existing procedures.

METHOD

The speech materials used in this study consisted of 12 spontaneous monologues (Dutch): 6 painting descriptions produced by 2 female speakers, MM and LK, amounting to 46.5 minutes of speech in total.

In a task that was individually performed, 38 subjects were instructed to mark paragraph boundaries in transcriptions of the monologues presented without interpunction or specific layout to indicate paragraph structure. Subjects were told to draw a line between the word that ended one paragraph and the one that started the next paragraph. No explicit definition of a paragraph was given. There were two conditions: half of the subjects could listen to the actual speech (SP condition), whereas the other half could not (TA condition). The reason to have both these conditions was to gain insight into the added value of prosody.

A typical example of part of a text is given below, followed by a literal translation in English. The two digits between round brackets represent the boundary strenght estimates, computed as the proportion of subjects indicating that there was a break, for the SP and TA condition, respectively. For sake of presentation, boundaries of strength 0 in either of the two conditions are only given when there is a stronger break in the other condition.

het is echt een paard dat [uh] over iets heel springt heel heel snel **(0.26; 0.11)** de man die d'r opzit die zit ook helemaal in zo'n gebogen [uh] [uh] ruitershouding met zijn billen omhoog en zijn [uh] hoofd **(0;0.05)** in de manen van het paard **(0.95;0.16)** het paard is wit **(0.11;0)** [uh] ruiter is is [uh] rozig rood **(0.53;0.79)**

(it is really a horse that [uh] jumps across something very very fast the man who sits on it he really sits also in such a bent over [uh] [uh] rider's position with his bum in the air and his [uh] head stuck in the mane of the horse the horse is white [uh] rider is is [uh] pinkish red)

As can be seen, the breaks between word clusters may vary between relatively weak ones (e.g., 0.05) to relatively strong ones (0.95). In total, the two monologues gave 889 'minimal units', i.e., sequences of words not seperated by any of the labelers.

RESULTS

Comparing SP with TA

A rough idea of the differences for the two conditions can be derived from figure 1, which shows the boundary strenght values for one typical monologue. The figure shows that the two

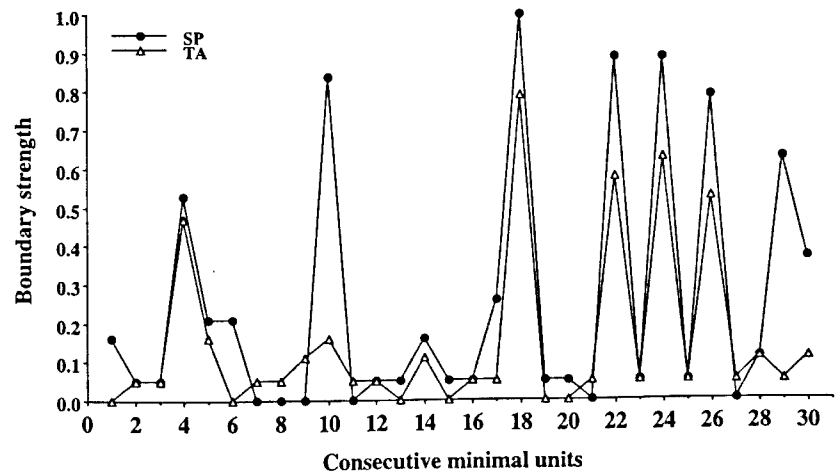


Figure 1: Boundary strength values for consecutive minimal units in SP and TA condition (further explanations in text)

experimental conditions give similar results, but segmentation is clearer in the text-with-speech case. The stronger breaks are more fully pronounced in the SP condition in the sense that proportionally more subjects agree on a paragraph transition. Also, although not visible in figure 1, some passages in the monologue receive different segmentations (indicating structural ambiguity) in the text-alone condition; such sections appear to be unambiguous when subjects have access to speech. This shows that prosody can limit the text interpretation.

Phonetic analyses

Given these observations, the speech was phonetically analyzed to explore potential relationships between boundaries obtained by the Rotondo method and prosodic features. Only the phonetic correlates of the boundaries in the SP condition are studied. Inspired by the literature, measurements include pitch range and pause, taking F0 maximum and silent interval as the respective acoustic correlates. The distribution of different boundary tones was also investigated. To have a unit of analysis, the monologues were transcribed by an independent labeler, who had to mark both the boundaries of phrases plus the sort of boundary tone at their respective ends (see below).

Pitch range In any given phrase, the highest F0 peak in an accented syllable was taken as a measure of pitch range ([2]). The results are given in Table 1 with the median values for pitch range in phrases following a boundary of a particular strength. The strength estimates are clustered into 5 groups, i.e,. one cluster containing values for breaks on which up to 25% of the labelers agreed, the next cluster having agreements between 25 and 50%, etc. Phrases within minimal units, i.e., units not subdivided by any of the labelers, were taken as a seperate category, since they formed a relatively large group. The median, rather than the mean, was taken as a more conservative measure to obtain a rough estimate, since the data in the different clusters were not always normally distributed. (A similar procedure is followed in Tables 2 and 3.) From Table 1 it can be seen that pitch range covaries with the depth of a discourse break. Pitch tends to become higher when the phrase follows a stronger boundary. This is true not only as an overall result, but also for the two speakers seperately.

Table 1: Median values of pitch range (in Hz) for speakers LK, MM seperately and pooled as a function of the different clusters of boundary strength

	0	0-.25	.25-.50	.50-.75	.75-1
LK	228	231	238	245	252
MM	221	231	239	238	245
Both	225	231	238	244	249

Pause Pauses were measured as silence intervals of at least 1 second long. Results are shown in Table 2, giving the median length of pauses preceding a boundary of a particular strength. It reveals a similar tendency as with the data for pitch range. Looking at the over-

Table 2: Median values of pause (in s) for speakers LK, MM seperately and pooled as a function of the different clusters of boundary strength

	0	0-.25	.25-.50	.50-.75	.75-1
LK	0	0	1.4	2.5	2.0
MM	0	0	1.9	1.6	4.5
Both	0	0	1.4	2.0	3.2

all data, it appears that pauses gradually become longer as a function of the strength of the discourse boundary, although the details for the two speakers seperately are somewhat more complex: the lengths for LK somewhat level off for the .75-1 cluster, whereas there is a sudden increase there for MM.

Boundary tone Finally, the distribution of different types of boundary tones

was investigated. Originally, the transcriber was instructed to mark phrases as ending in a low, mid or high boundary tone. In this paper, the latter two categories are collapsed into one, i.e., non-low boundary tones. Table 3, giving the pro-

Table 3: Proportion of low boundary tones for speakers LK, MM seperately and pooled as a function of the different clusters of boundary strength

	0	0-.25	.25-.50	.50-.75	.75-1
LK	.04	.18	.24	.27	.40
MM	.10	.29	.40	.38	.42
Both	.07	.22	.30	.31	.41

portion of low boundary tones, shows that the chance that such a tone will occur becomes higher as a function of the strength of the boundary.

DISCUSSION

The contribution of this paper is primarily methodological in that it presents a technique to analyse hierarchical discourse structure and its potential phonetic correlates in unrestricted discourse. It is a useful alternative to existing methods, as it is general and reproducible. A major disadvantage, however, is that the boundary strength measure (ideally) requires a large amount of subjects.

As for the prosodic results, it is interesting to see that prosodic variables such as pitch range, pause length and number of low boundary tones increase continuously with boundary strength at the discourse level. This is similar to prosodic phrasing results below the level of the sentence ([10]).

Of course, the features studied in this paper are not the only potentially interesting ones. In particular, preliminary observations suggest that transitions between major information units are accompanied by hesitation phenomena, such as filled pauses that point towards planning processes. These constitute an interesting area for further research. Also, future work will have to determine in what ways the experimentally based discourse boundaries correspond to junctures predicted by discourse theories.

ACKNOWLEDGMENTS

M. Swerts is also affiliated with the University of Antwerp (UIA). Thanks are due to R.-J. Beun for providing the speech materials, to E. Blaauw for transcribing them, and to R. Collier and A.A. Sanderman for commenting upon an earlier version of this paper.

REFERENCES

[1] Brown, G., Currie, K. and Kenworthy, J. (1980): *Questions of intonation*. London: Croom Helm.

[2] Grosz, B. and Hirschberg, J. (1992): "Some intonational characteristics of discourse structure," *ICSLP 92*, pp. 492-432.

[3] Lehiste, I. (1975): "The phonetic structure of paragraphs," *Structure and process in speech perception* (ed. by Nooteboom and Cohen), pp. 195-206.

[4] Passonneau, R.J. and Litman, D.J. (1993): "Intention-based segmentation: human reliability and correlation with linguistic cues," *ACL-93*.

[5] Rotondo, J.A. (1984): "Clustering analyses of subjective partitions of text," *Discourse Processes* 7, pp. 69-88.

[6] Sluijter, A. and Terken, J. (1994): "Beyond sentence prosody: Paragraph intonation in Dutch," *Phonetica* 50, pp. 180-188.

[7] Swerts, M. and Geluykens, R. (1994): "Prosody as a marker of information flow in spoken discourse," *Language and Speech* 37, pp. 21-43.

[8] Terken, J. (1984): "The distribution of pitch accents in instructions as a function of discourse structure," *Language and Speech* 27, pp. 269-289.

[9] Thorsen, N.G. (1985): "Intonation and text in Standard Danish," *JASA* 77, pp. 1205-1216.

[10] Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P.J. (1992): "Segmental durations in the vicinity of prosodic phrase boundaries," *JASA* 91 (3), pp. 1707-1717.

CONSISTENCY OF PROSODIC TRANSCRIPTIONS LABELLING EXPERIMENTS WITH TRAINED AND UNTRAINED TRANSCRIBERS

Matthias Reyelt
Institut für Nachrichtentechnik, Technische Universität Braunschweig, Germany
email: m.reyelt@tu-bs.de

ABSTRACT

For the use in the German Project VERBMOBIL a labelling system has been designed, that can be used by different project partners for a variety of purposes. It is based on the ToBI system for English and has some extensions to satisfy the special needs of individual project partners.

As the prosodic transcriptions have to be made by several transcribers having only little training, the achievable consistency was examined in several experiments.

Two main labelling experiments are described here: the first, with fully untrained transcribers in order to obtain a starting value for inter transcriber consistency; the second, after a training phase to examine the improvement in consistency achieved by the training.

INTRODUCTION

Although prosody has been investigated for several decades, the resulting knowledge has rarely found its way into automatic speech recognition. One reason for this might be that the statistical methods like HMM and statistical grammars, that seem to be a current standard in speech recognition, need large amounts of labelled speech data for training to produce reliable results.

However, for the recognition of spontaneous speech also prosodic information is needed. One of the aims of the German project VERBMOBIL [5] is the integration of prosodic information at all levels of the recognition process. The prosodically labelled data needed for training and test are produced centrally for all project partners.

There are several demands made on such a system for prosodic labelling. For a data driven training of speech recognition systems the amount of labelled data must match the number of different labels. If a very detailed inventory is used, a lot of speech material has to be labelled before the data are of any use for automatic speech recognition.

The inventory has to meet the different needs of different users. Such an inventory is always a compromise between people.

It is of great importance that the labelled data be ready for automatic analysis. Machine readability and formal consistency are indispensable. A lot of work in the SAM-Project was devoted to the development of label inventories and standard labelfile formats that also include prosodic information [1].

On the other hand there are requirements from the labelling point of view. Larger amounts of data have to be labelled by different transcribers after only a short training phase. Although some subjective variation might be inevitable, the inter-transcriber consistency has to minimized by carrying out several measures:

The label inventory has to be transparent, i.e. it has to match to the perception of transcribers who do not have a profound knowledge of prosodic theory.

A permanent evaluation is necessary to keep track of weaknesses of the system.

A system for labelling large corpora is the ToBI system for English prosody [4]. This system was developped in accordance with the above criteria and has become (or is on its way to becoming) a standard system for transcribing prosody. The labelling system described here is an adaptation of this system for German prosody.

DESCRIPTION OF THE LABELLING SYSTEM

The labelling system used in these experiments is divided into three tiers, which are partially similar to the ToBI-system:

The functional tier

In this tier a more "functional" prosodic labelling is performed. The tier is not part of the original ToBI and is therefore described in detail:

One part of this tier is the labelling of *sentence modality*. This might not be part of a core prosodic analysis but is clearly suprasegmental and is needed by several project partners.

The other part is the basic labelling of *accented words* based on auditive impression. There are three different accent types: *secondary accent*, *main accent* and *emphatic/contrastive accent*. In each intonational phrase the most prominent word obtains the *main accent*[1]. Although this is of course not a focus analysis, it offers some information about the focal structure of the utterance.

There are several reasons for introducing this additional tier:

- It is a customer's tier. The information in this tier was needed by partners.
- Together with the break index tier it represents a basic system that can be labelled faster and with less training than a "full" labelling including the tone tier.
- An analysis of the labelled data showed that the syllable durations correspond to the accent type. This tier seems to hold additional information about accents that is not labelled in the tone tier.

The tone tier

In this tier *pitch accents*, *phrase accents* and *boundary tones* are labelled using an inventory similar to ToBI.

The break index tier

This tier, too, is quite similar to the break index tier in ToBI with slight formal changes in the index numbering: *intermediate phrase boundary* (B2), *intonational phrase boundary* (B3) and *irregular boundary* (B9).

EXPERIMENTS

Using this inventory, labelling experiments were carried out. Several subjects made parallel transcriptions of the same material.

In a first experiment [2] [3] five subjects labelled 480 utterances of the PHONDAT92 corpus[2]. The subjects had no experience and only a short introduction to their task. Only the functional tier and a reduced break index tier were used. The transcriptions were based merely on auditive perception, no visual aids such as F_0-contour were given.

After this experiment a training programme was developped and in a second labelling experiment the tonal tier

[0]This work was funded by the German Ministry of Education, Science, Research and Technology, Contract No. 01IV101N0. The responsibility for the contents lies with the author

[1]This is not a strict rule; where appropriate, there can be more than one *main accent* per phrase. The *main accent* can be replaced by an *emphatic accent*.

[2]The PHONDAT92 corpus consists of single read utterances from a travel inquiry scenario.

was included as well[3]. For the second experiment 233 utterances from the VERBMOBIL corpus[4] were used.

LABELLING ENVIRONMENT

The labelling was carried out on a workstation using *fish*, a labelling software based on Tcl/Tk, that is easy configurable and supports the SAM format for labelfiles.

In the first experiment only the speech signal and the orthographic text was displayed, in the second experiment the pitch contour was added.

STATISTIC EVALUATION

In the first experiment the subjects labelled 480 utterances. The resulting 5520 pairs gave an overall correspondence[5] of 80% for the accents (*secondary* and *main accent*) and 94% for *phrase boundaries* (no further distinctions).

However, this overall correspondence is only a rough evaluation. Additionally the distributions of accent and boundary types are rather unequal and the unaccented syllables make a major contribution to the value.

Thus an independent evaluation value was calculated for each accent/boundary class according to equation 1.

Equation 1: Calculation of label dependent correspondence $corr_{1,2,label}$. $n_{corr(1,2),label}$ *is the number of correct pairs for a particular label.* $n_{1,label}$ *and* $n_{2,label}$ *are the total numbers of this label occurring in each of the transcriptions*

$$corr_{1,2,label} = \frac{n_{corr(1,2),label}}{(n_{1,label} + n_{2,label})/2} \quad (1)$$

This leads to the correspondence values shown in Table 1:

Table 1: Inter–transcriber correspondence reached by untrained transcribers in the first experiment

secondary accent	40 %
main accent	72 %
phrase boundary	76 %

The percentages in Table 1 show a satisfying correspondence for *main accent* and *phrase boundary*. For *secondary accent* the correspondence is much lower and shows the transcribers' uncertainty in the decision *accented/unaccented*.

In the second experiment the subjects had a training phase with a number of selected utterances to introduce the label inventory and then a nine-dialogue experience. For the evaluation five different dialogues were chosen, consisting of 233 utterances (2907 pairs). The overall inter–transcriber correspondence is listed in Table 2.

Table 2: Overall correspondence in second experiment

functional tier	91 %
break index tier	94 %
tone tier (pitch acc.)	85 %
tone tier (boundaries)	88 %

Again the correspondences for the individual labels were calculated. Table 3 shows the results for the functional tier and the break index tier. For the tone tier the correspondence varied widely, from maxima of about 56% for H* and L*+H pitch accents down to an absolute minimum of zero for the downstepped L*+!H accent (which occurred only four times). For boundary tones the max. correspondence was 75% for the L-L% boundary, the minimum was 35% for the L-H% boundary.

[3] Unfortunately only two of the five subjects remained from the first experiment (it seems indeed that prosodic labelling is not that much fun for most people, why?), so the results of this experiment remain preliminary.

[4] The VERBMOBIL corpus consists of spontaneous negotiation dialogues.

[5] This correspondence is calculated according to the ToBI system, see [4]

Table 3: correspondences for individual labels, second experiment

secondary accent	32 %
main accent	86 %
intermed. phrase bound.	44 %
intonational phrase bound.	90 %

The correspondence values are better than in the first experiment, at least for *main accent* and *intonational phrase boundary*. For the *secondary accent* the correspondence has decreased; the distinction accented/unaccented is still rather uncertain.

ANALYSIS OF THE TRANSCRIPTIONS

The statistical evaluation gives an overview over the consistency between the transcribers. However it provides no information about the reasons for the different transcriptions and may even hide errors if they are consistently made by all transcribers.

Additionally a more profound analysis of the transcriptions is necessary in order to examine errors and misinterpretations of the labelling system. Such an analysis showed a variety of reasons for differing transcriptions.

Especially the first experiment revealed that consistency is speaker dependent to a high degree. The quality depends on how familiar the transcriber is with the speaker's dialect. Besides, the label inventory and the training do not (yet) cover all German dialects and speaking styles.

Different transcriptions are also caused for several other reasons. Firstly, the categorial boundaries between the labels (e.g. H* and L+H*) are not always clearly distinguishable. Secondly, misinterpretations of the pitch contour lead to erroneous transcriptions. Thirdly, the usage of particular labels was misunderstood by the transcribers.

In an additional training (in particular using erroneous utterances) the number of labelling mistakes can surely be reduced. However, a regular consistency check seems to remain necessary.

OUTLOOK

Although these experiments are preliminary, they provided useful insights into practical problems of prosodic labelling. As a result, the training programme has been extended to include the difficult cases.

Moreover the labelling environment has been extended by providing means for the transcribers to mark their uncertainties and to add comments on their transcriptions.

The current database consists of approx. one hour of labelled speech that has already successfully been used by several project partners.

REFERENCES

[1] *User guide to ETR tools.* SAM-UCL-G007, pp. 15–19, 1992.

[2] Matthias Reyelt. Experimental investigation on the perceptual consistency and the automatic recognition of prosodic units in spoken German. In *Working papers*, volume 41, pages 238–241, Lund University, 1993. Dept. of Linguistics.

[3] Matthias Reyelt. Untersuchungen zur Konsistenz prosodischer Etikettierungen. In H. Trost, editor, *KONVENS 94*, pages 290–299, Berlin, 1994. Springer.

[4] Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. Tobi: A standard for labeling english prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, pages 867–870, 1992.

[5] W. Wahlster. Verbmobil: Translation of face-to-face dialogs. In *Proceedings Eurospeech 93*, 1993.

TONES IN MPUR (West Papuan Phylum)*

Cecilia Odé
Irian Jaya Studies, Leiden University, The Netherlands

ABSTRACT

An experiment has been carried out in order to verify the number and types of lexical tone contrasts in *Mpur*. Words in presumed minimal tone pairs were presented to native listeners in original and manipulated versions. The tasks were to translate the words from *Mpur* into Indonesian and back. The results justify classifying *Mpur* as a language with lexical tone contrasts.

BACKGROUND

Mpur (West Papuan Phylum, Bird's Head Superstock, Amberbaken Stock-level isolate (Voorhoeve 1994:73ff) [1]), in the literature often referred to as *Kebar* or *Amberbaken*, is spoken by ca. 5000 speakers in the Kebar Valley and Amberbaken (northeast Bird's Head, Irian Jaya) and one of the few languages with lexical tone contrasts in the area. It has two dialects: Sirir (on the coast) and Ajiw (in the mountains) (cf. Kalmbacher 1990:2 [2]).

THE QUESTION OF TONE

During fieldwork (1993/94) I recorded spontaneous texts (myths of origin, folk tales, daily life stories), a vocabulary (2000 entries) and some prepared texts. Evidence for phonemic tone was easily found in word strings with tonal opposition, but number and types of tone had to be specified yet (cf. Kalmbacher 1990:18ff [2]). Instead of impressionistic data described from hearing, an experimental phonetic approach of the issue, in which presumed types of tone are verified in perception experiments with native listeners, would enable me to give a description of types of phonemic tone with phonetic specifications. Experimentally verified data of tone, expressed in relative values (see below), will then be fully accessible and can be reproduced. I conducted an experiment in which words, classified into perceptually and phonetically similar types of pitch level and pitch movement, were presented to native listeners in original and manipulated versions. For the classification I selected 119 words in isolation and 112 words in a small context, pronounced by one female and two male native speakers after an Indonesian translation. The corpus consisted of 91 monosyllabic, 105 disyllabic and 35 trisyllabic words; a total of 406 syllables. The selection of words was made on the basis of two criteria: i) the occurrence of a word in minimal pairs (70 words), triplets (30 words), quadruplets (16 words) and quintuplets (5 words); ii) the types of pitch level or movement in mono-, di- and trisyllabic words, which did not occur in word strings with tonal oppositions (110 words). The classification procedure was as follows. I stylized pitch movements in all 231 words by means of the analysis-by-resynthesis stylization method developed at IPO (Eindhoven, The Netherlands), described in 't Hart et al. (1990) [3]. In this method, measured fundamental frequency (F0) curves (Hermes 1988 [4]) are replaced by the smallest number of straight-line segments which still yield perceptual equality with the original F0 curves. The stylized fragment, represented on a logarithmic scale in semitones, can be made audible and compared with the original F0 curve of the same fragment; no differences may be audible. If pitch in a syllable could be stylized into a level straight-line segment without audible difference with the original pitch, it was defined as a *level* pitch; otherwise pitch was defined as a pitch *movement*. If native listeners consistently distinguish between these types of pitch level and pitch movement, in a phonemic representation types can be defined as level tones and contour tones, respectively. On the basis of the stylizations I classified syllables that were phonetically similar into types of pitch level or pitch movement. The relative values of pitch in each syllable in semitones enabled me to classify similar syllables pronounced by speakers with different pitch ranges. For the 406 syllables, the types I arrived at (the numbers per type are given between brackets) were:

- five types of pitch level, i.e. high (49), midhigh (26), mid (161), midlow (81), low (28) (henceforth: H, MH, M, ML, L) with a pitch range of ten semitones (ST) between high and low, and
- three types of pitch movement, i.e. mid-falling (14), low-falling (5), falling-rising (42) (henceforth MF, LF, FR) with an excursion size of 5 ST in each movement.

The question was, whether the eight phonetic types are phonemic: five level, three contour tones. Such a complicated tonal system was unlikely to exist: the interval between two nearest pitch levels seemed very small (ca. 2 ST) for tonal contrasts. Moreover, realizations with *both* level types H and MH, or MH and M in the *same* word occurred. The same holds true for level types M and ML, ML and L, and for movements MF and LF. Therefore, in the stimuli for the experiment I reduced the five types of pitch level to three (H, M, L), changing type MH into H and/or M and type ML into M and/or L with an interval between two levels of ca. 5 ST. Furthermore, the question was whether types MF and LF are two or just one phonemic tone, or whether they are contextual variants of pitch level MH/M/ML and ML/L, respectively, the movements being ascribed to the interaction of (inherent features of) syllable-final consonants/vowels with tone or intonation. In the stimuli for the experiment I changed types MF into M, L and LF, LF into M, L and MF. Finally, the question was whether type FR is phonemic and realized with two pitch movements within one syllable. Summarizing, questions were:

1) How many types of pitch level are phonemic?
2) Are types MF, LF one or two contour or one or two level tones?
3) Is type FR a complex contour tone?

The experiment was carried out during fieldwork in February '95.

THE EXPERIMENT

The experiment consisted of a *perception* and a *production* task with 146 stimuli: 51 mono-, di- and trisyllabic words in the *original* realization and 95 *manipulated* versions (103 manipulated syllables) of these words. The 51 *original* stimuli were selected from the 231 words dicussed above, pronounced by one of the male speakers: a literate, thirty years old son of a Kebar mother and an Amberbaken father. In order to avoid confusion if more speakers with different pitch ranges were used, and to avoid introducing dialectal variants, the selection of one, in my opinion very consistent, speaker seemed justified. His phonetic specifications (mean values for each type), according to which manipulations were made, are: H 190 hz; MH 170 hz; M 145 hz; ML 125 hz; L 100 hz; MF 140-100 hz; LF 120-90 hz; FR 145-100-145 hz.

Stimuli for the first two questions are:

- eight monosyllabic tokens in six minimal pairs and two minimal triplets, with two or four manipulations: 18 original, 42 manipulated = 60 stimuli;
- seven disyllabic tokens in minimal pairs with one manipulation: 14 original, 14 manipulated = 28 stimuli;
- seven di- and two trisyllabic words (*not* occurring in minimal pairs) with one manipulation: 9 original, 9 manipulated = 18 stimuli.

Stimuli for the third question:

- one mono- and nine disyllabic words (*not* occurring in minimal pairs) with three manipulations: 10 original, 30 manipulated = 40 stimuli. For numbers and types of manipulations see the results. All stimuli were shuffled and randomly recorded on tape in a *resynthesized* version. Practice stimuli preceded the tasks described below.

The *perception* task listeners had to perform was to listen to the *Mpur* words recorded on tape, to give a translation into Indonesian of only correctly pronounced words which was written down by me, and to give no translation if a word was pronounced incorrectly or was unrecognizable.

The *production* task was performed a few days later and consisted of a translation into *Mpur* of the Indonesian words (as they were given last year for the *original* recordings, i.e. *not* the translation by listeners in this experiment), read aloud and recorded on tape.

I suggested, that if the translation into Indonesian of the original and the manipulated versions of one word was the same, the original word had been recognized consistently and the manipulation had been executed correctly. If a different or no Indonesian translation was given, the realization of pitch level or pitch movement in the *production* task must also differ. This could be verified by measuring and comparing the original and the new realization of the stimulus. I expected that ultimately three pitch levels (types H, M, L) and two pitch movements (types LF, FR) would be found to exist; my manipulations would be correct then.

Two trained native listeners, one male (the speaker of the stimuli, see above) and one female (forty years old) performed the tasks. Other listeners invited were all non-trained, and the tasks proved to be too difficult. But, as we will see, the present two listeners were consistent in their judgements.

RESULTS AND CONCLUSION

In the *perception* task of the experiment, the *original* versions of 49 stimuli out of 51 were translated according to the original translation. The two exceptions were stimuli */ip/* (wind) and */wot/* (to see) realized with type MH, which in the manipulated version were accepted as type M, but rejected as type H, because of a tonal contrast in the same pair of type H vs. M. In the *production* task, the translation from Indonesian into *Mpur*, the same 49 stimuli were realized with the same type as in the original recording; the two exceptions were now realized with type M. I verified the phonetic similarity of each stimulus pair, i.e. the original and the new realization, by measuring and comparing pitch in both versions, using speech analysis system "Cecil" version 2.0, developed by Jaars Inc. USA (it had to be done in the field). No dissimilarities were found to exist, which was confirmed by the listeners. After the official experiment they compared the two versions presented in pairs via Cecil and accepted them as perceptually equivalent.

For the 95 *manipulated* versions (103 manipulated syllables) the results are presented below. The column "nr." indicates the number of manipulated syllables, the columns "yes" and "no" whether a given manipulated type was accepted or not. The results are not differentiated per listener, since they agreed in their judgements, except for two manipulations. I decided for the accepted version. Note, that the *monosyllabic* stimuli were manipulated into *two* or *four* types *each*.

Monosyllabic stimuli in minimal pairs and triplets:

type	nr.	yes	no
8 MH → H	8	5	3
M	8	5	3
6 ML → M	6	3	3
L	6	5	1
2 MF → M	2		2
L	2		2
MF	2	2	
LF	2	2	
1 LF → M	1		1
L	1		1
MF	1		1
LF	1	1	
1 FR → MF	1		1
LR	1		1
total	42	23	19

Type MF → MF and LF → LF are stylizations. Types H and M for MH, and types M and L for ML were two times *both* accepted; there was no tonal contrast in the same pair or triplets of types H vs. M, or M vs. L.

Disyllabic stimuli in minimal pairs:

type	nr.	yes
MH → H	1	1
MH → M	2	2
ML → M	6	6
ML → L	9	9
MF → M	1	1
LF → L	1	1
total	20	20

Di-, trisyllabic stimuli not in pairs:

type	nr.	yes	no
MH → H	3	2	1
ML → M	2	2	
ML → L	6	4	2
total	11	8	3

The three rejected manipulations occurred in one word; according to the listeners, type H was too high; type M would be acceptable.

Stimuli of type FR:

type	nr.	yes	no
M-FR →			
M-MF	9	2	7
MF-LR	9	1	8
L-LR	9		9
FR → MF	1		1
→ LR	2		2
total	30	3	27

In the *production* task of words with type FR, the two words for which type M-MF and one word for which MF-LR was accepted, were realized with type M-FR. I have no explanation for accepting the types here.

The three questions, formulated above, can now be answered.

1) *Number and types of pitch level.* The results show, that types MH and ML are not phonemic: if in a given word string an opposition exists of type H vs. M, type MH is a contextual variant of either type H or type M; if in a given word string there is no opposition of type H vs. M, the type can be realized as H, MH or M. The same holds true for type ML. For example, type MH in */muk/* (tail) was accepted with types H and M, type ML in */muk/* (name) only with type L; type ML in */pa/* (already) was accepted with types M and L, type MH in */pa/* (rain) only with type H.

2) *Types MF and LF.* There are only three examples of these types, since in a lot of stimuli falling movements were stylized into level tones and accepted (see above). Type LF in */bak/* (axe) is the only acceptable realization, and for type MF in */ip/* (boil) and */džan/* (not) both MF and LF were accepted; for all three stimuli level types were rejected. Afterwards, listening to type MF and LF stimuli again, the native speakers were persistent in their judgement and came up with more examples of type LF. For the time being, I accept contour tone LF, since manipulations of type MF into LF were acceptable.

3) *Type FR.* The results show that this type is only accepted in its original realization. Other realizations are incorrect or a dialectal variant (type LR): both listeners confirmed my earlier observation, that in the given words type LR is regular in the Ajiw dialect.

Finally, tone contrasts are presented below. They are marked with +, but tone contrasts only occurring in final syllables of polysyllabic words are marked with x. Note, that types LF and FR were not observed in initial or central syllables of polysyllabic words.

Tone contrasts:

	H	M	L	LF	FR
H		+	+	+	
M	+		+	+	+
L	+	+		+	+
LF	+	+	+		x
FR		+	+	x	

ACKNOWLEDGEMENT

Imbwar saswar na iniman Wasyaben Ajo braw inimata Kombwara Wabya ma dobot batut wasi Mpur braw dorin bar doritot braw in. (I thank my friends Wasyaben Ajoi and Kombwar Wabia for giving information about the *Mpur* language and for being and laughing with me.)

* This research is part of ISIR (Irian Jaya Studies), A Priority Programme supported by the Netherlands Organization for Scientific Research (NWO).

REFERENCES

[1] Voorhoeve, C.L. (1994), *Comparative Linguistics and the West Papuan Phylum*, Maluku & Irian Jaya, E.K.M. Masinambow ed., LIPI Jakarta, pp. 65-90.

[2] Kalmbacher, J.G. (1990), *Mpur phonology*, Cenderawasih University & SIL, unpublished ms., Jayapura 1990.

[3] Hart, J. 't, Collier, R., Cohen, A. (1990), *A perceptual study of intonation: An experimental phonetic approach to speech melody*, Cambridge.

[4] Hermes, D.J. (1988), "Measurements of pitch by subharmonic summation", *J.Acoust.Soc.Am.*

TONAL ALIGNMENT AND THE REPRESENTATION OF ACCENTUAL TARGETS

Amalia Arvaniti & D. Robert Ladd

Department of Linguistics, University of Edinburgh, U.K.

ABSTRACT

This paper examines the tonal composition and alignment of prenuclear accents in Greek. Our experimental results suggest that these accents, which show an initial dip and late peak alignment, are best described as L*+H since (a) their initial L tone is invariant in scaling and alignment and not affected by declination, and (b) the H tone is more variable in alignment and affected by the position of the accented syllable within the word.

1. INTRODUCTION

Prenuclear pitch accents in Greek show a slow rise that begins on the accented syllable and reaches its peak towards the end of this syllable, or on the following one (see Figure 1 for an example).

In the standard autosegmental-metrical framework of intonational analysis, as exemplified by [4], such accents in English are described as L*+H and said to be different from the two other types of rising accent, H* and L+H*, in both scaling and alignment: H* does not show the bitonal accents' initial dip, while the difference between L*+H and L+H* relates to the variable alignment of the "unstarred" tone (the trailing H and the leading L respectively). The evidence, however, for the three accent types has been disputed, e.g. by Ladd [3], who has argued that all rising accents are instances of H* with variable peak alignment, not distinct categories.

This issue is still unresolved in English. However, if it could be shown that another language, in the present case Greek, uses at least one of the bitonal accents, then the necessity of differentiating between single and bitonal rising accents in the universal inventory of accent types would have been demonstrated.

2. METHOD

In order to examine the tonal composition of Greek prenuclear accents, and in particular whether the L tone needs to be specified in their phonological representation, we devised two sets of sentences in which two accents (A1 and A2) within the same intonational phrase were separated by progressively more unaccented syllables. Table 1 gives the details of one of the sets of sentences. The second set was constructed along similar lines, but gave less useful results because speakers tended to divide some of the sentences into two prosodic phrases. This is discussed further below.

The hypothesis was as follows: if the F_0 dip observed at the beginning of A2 is due to declination between two H* accents, then it would become deeper as the number of unaccented syllables between A1 and A2 increased; if the F_0 dip is due to the specified L tone of a bitonal accent, the alignment and scaling of this tone would remain relatively stable regardless of the number of unaccented syllables between A1 and A2.

The test sentences were recorded in a sound treated booth in the Phonetics Laboratory of the University of Oxford. Three native speakers of standard Greek, naive as to the purposes of the experiment, recorded seven repetitions of the sentences of both sets, in random order. Durational and F_0 measurements of the four most natural (in the first author's judgement) repetitions were obtained from waveforms and F_0 traces respectively, using Waves+. The F_0 measurements were transformed into ERB scale (see [1], [2]).

***Table 1**: One of the two sets of test sentences. The syllables bearing A1 and A2 are underlined.*

1. [tilefo'no sto 'mano ja to 'parti]
   "I'm calling Mano about the party."
2. [tilefo'nusa sto 'mano ja to 'parti]
   "I was calling...."
3. [tilefo'nusame sto 'mano ja to 'parti]
   "We were calling...."
4. [tilefo'nusame me to 'mano sti 'meri ja to 'parti]
   "We and Mano were calling Mary...."
5. [tilefo'nusame apo to 'mano sti 'meri ja to 'parti]
   "We were calling from Mano's to Mary...."

The data were analysed statistically using analyses of variance in which the independent variables were speaker and number of unaccented syllables between accents. Where necessary, the ANOVAs were followed by Scheffé tests; *p*-levels for these are presented below.

3. RESULTS AND DISCUSSION

The results on scaling show that for speakers CN and NP the value of the L tone of A2 (L2) is not affected by the number of unaccented syllables between accents (see Figure 2). ET's data, however, show a weak effect of this factor: the value of L2 is higher for sentence 1, with one unaccented syllable between accents, than for sentence 4, with four unaccented syllables (p = 0.03).

In CN and NP's data the number of unaccented syllables did not affect the F_0 difference between L2 and the H tone of A1 (H1). In the data from ET, however, this difference increases as unaccented syllables are added: it is smaller for sentence 1 than for sentences 3, 4 and 5 (p = 0.018, p = 0.001 and p = 0.0001 respectively); it is also smaller for sentence 2 than for sentence 5 (p = 0.02).

In terms of alignment, L2 is consistently aligned with the beginning of the stressed syllable bearing A2, in the data from all speakers (see Figure 3).

In contrast, the alignment of both H1 and H2 (the H tone of A2) exhibits greater inter- and intra-speaker variability as Figure 3 shows. In the case of H1 in particular, the peak is reached further from the beginning of the accented syllable as the number of unaccented syllables increases (for speakers ET and CN): H1 is closer to the beginning of the accented syllable in sentence 1 than in sentences 3, 4 and 5 (for ET, p = 0.009, p = 0.007 and p = 0.001 respectively; for CN, p = 0.0001 in all cases).

As these differences in alignment level off once the number of three unaccented syllables is reached, i.e. once the accent is placed on the antepenult (see Figure 4), the results suggest that the alignment of this H tone may depend on the position of the accented syllable relative to the right boundary of the accented word.

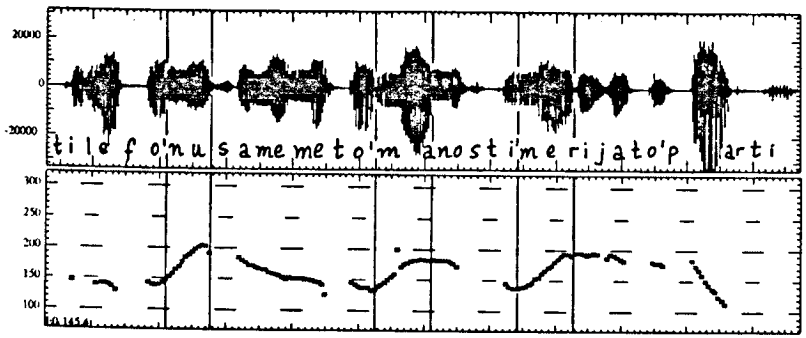


***Figure 1**: Waveform and F_0 contour of the test sentence [tilefo'nusame me to 'mano sti 'meri ja to 'parti]. The parts of the contour corresponding to prenuclear accents are between vertical lines.*

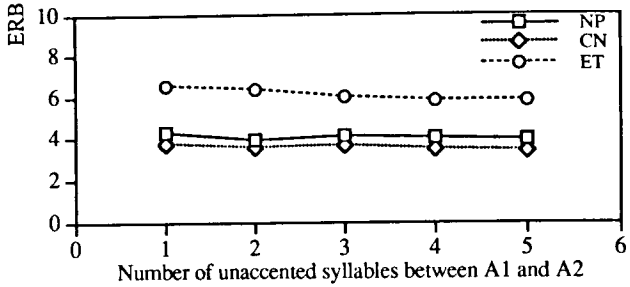


Figure 2: Scaling of L2 as a function of the number of unaccented syllables between accents, for each speaker separately. Mean values are in ERB scale (standard deviations are too low to be shown).

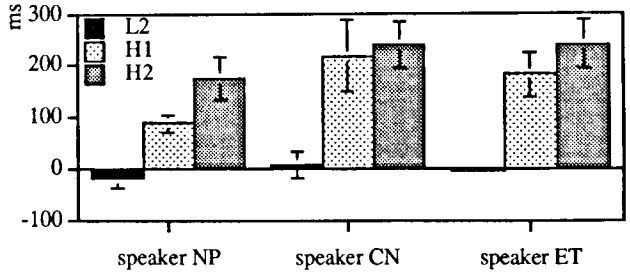


Figure 3: Alignment of L2, H1 and H2, i.e. distance of the tone from the beginning of the accented syllable (in ms), for each speaker separately; means and standard deviations are shown.

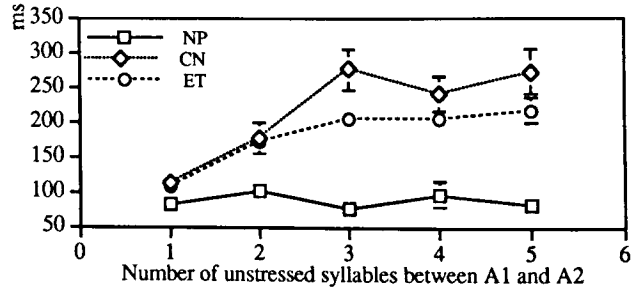


Figure 4: Alignment of H1, i.e. distance of the tone from the beginning of the accented syllable (in ms), as a function of the number of unaccented syllables between accents, for each speaker separately; means and standard deviations are shown.

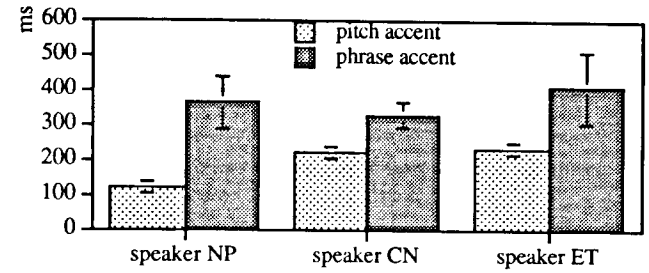


Figure 5: Mean distances between L2 and H2 (and standard deviations) in the second set of sentences, for each speaker separately; light grey bars show the distance when the L and H tones form a bitonal accent, and dark grey bars when they form the sequence L H⁻ (see penultimate paragraph of text).*

Thus the data suggest that prenuclear pitch accents in Greek are best represented as L*+H, since the L tone (a) is clearly specified and not a result of declination and (b) shows more stable scaling and alignment than the H tone.

It might be argued that the results of the H tone alignment – viz. the fact that the H tone seems to align with the edge of the accented word – could justify considering it a type of phrase accent that demarcates the end of the word. However, the data from the *second* set of test sentences suggest that this is not an appropriate interpretation. The data from this set, as noted above, could not all be used in the main analysis, because in many cases speakers divided the sentences into two prosodic phrases. (In the cases where speakers did *not* divide the sentence into two phrases, the data from the second set agreed with the first set.) In autosegmental terms, when speakers divided the sentence into two prosodic phrases, they inserted a H⁻ phrase accent after the word which was intended to have A2, and replaced this accent with L*. In these cases, the alignment of the L and H is markedly different from the same tones in the L*+H accent ($F(1, 2) = 19.09$, $p = 0.04$, for the distance between L2 and H2 in the two configurations). As can be seen in Figure 5, when the two tones form a bitonal L*+H accent the distance between them is shorter and varies less; when the H tone is in fact a H⁻ then it is placed further away from the preceding L* accent, and the distance between the two tones is highly variable. In other words, the rise associated with prenuclear accents in Greek must be the trailing tone of a bitonal accent.

In conclusion, although further research on the alignment of the trailing tone of these accents and on the function and alignment of single H* accents is still necessary, the present results suggest that the prenuclear accents of Greek are best represented as bitonal L*+H accents.

REFERENCES

[1] Glasberg, B. R. & B. C. J. Moore (1990) Derivation of auditory filter shapes from notched-noise data. *Hearing Research* 47: 103-138.

[2] Hermes, D. & J. van Gestel (1991) The frequency scale of speech intonation. *Journal of the Acoustical Society of America* 90: 97-102.

[3] Ladd, D. R. (1983) Phonological features of intonational peaks. *Language* 59: 721-759.

[4] Pierrehumbert, J. (1980) *The phonology and phonetics of English intonation.* Ph.D. dissertation, MIT.

EVALUATION OF AUTOMATIC GENERATION OF PROSODY WITH A SUPERPOSITION MODEL

Y. Morlec, V. Auberge and G. Bailly
Institut de la Communication Parlée, INPG & Université Stendhal
46, av. Félix Viallet 38031 Grenoble Cedex 01, France
e-mail: (morlec, auberge, bailly)@icp.grenet.fr

ABSTRACT

A new paradigm for modelling prosody is introduced. We assume that global melodic prototypes are built and stored in a "prosodic lexicon". The actual generation of adequate prosodic contours is achieved by retrieving and combining these elementary global contours accessed by linguistic keys. Two automatic F0 generation procedures have been used: The first consists of a structured lexicon, the second uses a recurrent neural network. Preliminary results show that both methods provide F0 contours which can compete with natural ones.

THEORETICAL FRAMEWORK

The intonation of an utterance is classically described in terms of tone units regarded as the primary units of intonational structure [9] [10]. So-called pitch targets are the phonetic realisations of a limited set of phonologically distinct tone segments, typically less than ten. The dynamics of tones is often constrained by an utterance template consisting of upper-lines and base-lines.

Within this framework the structural coherence of pitch movements is ensured by higher phonological components. Our approach aims to associate these higher phonological units more directly with their prosodic instanciations via a superposition model. For each phonological level, global prosodic movements achieve the necessary contrasts: phonologically-relevant information is thus distributed and enables priming [7]. The prototypic prosodic movements signal level-specific contrasts such as modality of the utterance within the discourse or strength of linguistic boundaries between groups of words within the utterance.

The actual prosodic contour results from a superposition of prototypes where upper-level ones are minimally anchored

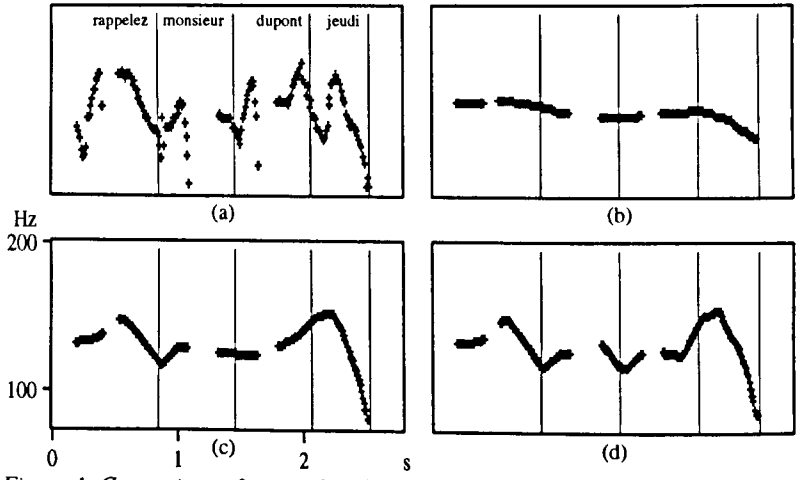


Figure 1. Comparison of original and synthetic contours obtained by the lexicon approach for the sentence: "Rappeler Mr Dupont Jeudi ! ". (a) the original curve, then successive superposition of (b) the sentence, (c) the syntagm and (d) the prosodic group. Note the lack of micromelody which could be easily produced with an additional level.

onto lower-level units.

In the next section, we test two superposition models: a simple additive model using a structured lexicon and a non-linear superposition model using a sequential neural network.

LEARNING PROCEDURE

Two automatic learning procedures have been applied to a corpus of 88 imperative utterances produced by one male speaker at a comfortable speech rate. The melodic curves were stylised with three values per vowel [5].

Three phonological levels directly related to the linguistic structure are considered here: (a) the sentence, (b) the syntagm, (c) the prosodic group (grouping each content word together with its function words).

The contours related to (b) and (c) are supposed to mark the degree of cohesion of the adjacent units as proposed in [2].

A structured lexicon

A methodology for developing a structured lexicon of prosodic contours has been already described in [1]. The data corpus is processed in a top-down hierarchy: upper-level prototypic contours are iteratively extracted by averaging and are then subtracted.

We developed a simplified version of this model which uses a normalised time-axis. A lexicon of prototypic melodic curves was built using a fixed 4th order polynomial interpolation where contours are scaled linearly to fit the considered linguistic boundaries. The figure 1.b shows the prototype for the sentence level. This prototype was then subtracted from the original contours and syntagmatic sub-contours are then grouped according to their relational marker and further processed.

The superposition model of generation then consists of a simple additive model (see Figure 1) which warps the prototypic melodic curves onto the actual syllables. Perceptual experiments described below show that this simple method leads to acceptable F0 generation. However, it is obviously too simple in its present form to adequately describe some important factors affecting melodic contours:

- It doesn't take account of the syllabic "weight" of the cued linguistic units.
- As f0 is the audible consequence of articulatory movements, undershoot can occur and thus speech rate, as well as stiffness of gestures, influence the actual realisations of intended targets.

Both restrictions mentioned above in addition to those imposed on the global shape of each melodic curve by polynomial interpolation could be solved by storing parameters of a dynamic equation. Sequential Neural Networks (SNNs) are known to model non-linear dynamics [8].

A neural network

In parallel with the structured lexicon, we performed simulations with SNNs. Although greatly inspired by the pioneering work of Traber [11], our approach differs in the characterisation of the input task: Traber uses a large window (13 symbols including syllables and phrase/word boundaries) on a linear phonological representation of the input sentence where major and minor accen-

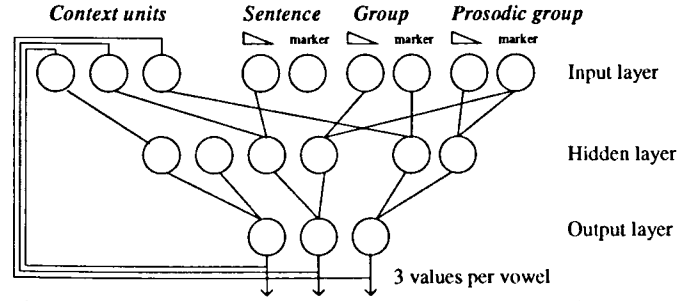


Figure 2. Structure of our SNN. Ramps signal the extent of associated units connected by limited sets of markers.

tual positions are already given.

Our network (see Figure 2) is responsible for transforming simple linear movements (ramps) into more complex contours according to level-specific strategies, and for superposing them in order to mimic the original melodic curve.

Instead of a large window on a phonological description as used by Traber, our network input consists of only two parameters per level (sentence, syntagm, and prosodic group) :

Prosodic markers :

They indicate how subsequent units of the same level are linked. Linguistic function of prosody is thus restricted here to signal relations entertained by multi-layered linguistic units.

Syllabic ramps :

They signal the "syllabic distance" from the current syllable to the next marker.

- The output consists of 3 values per vowel.

The network was trained with half of the corpus. The other utterances were kept for prediction tests. Encouraging results were obtained with this basic recurrent network. Experiments described below showed that it was even able to learn the systematic initial emphatic stress used by our speaker for this specific task.

PERCEPTUAL EVALUATION

Method

A preference test was designed to evaluate the perceptual relevance of these two methods : 10 triplets of sentences (giving 60 presented pairs) were generated using a high-quality TD-PSOLA analysis-resynthesis technique [3].

The A version is the natural utterance only degraded by our F0 description (3 points per vowel). The B and C configurations are obtained by the structured lexicon and the SNN respectively.

Seven subjects participated in this perception experiment. During the preference test, the subjects were asked to choose the most natural utterance from each presented pair.

Results

Considering all subjects, results show:

- When the A version is presented against B or C, it is identified only 70% of the time. This demonstrates that the models have captured essential features of the original prosodic contours.

- The C version is only occasionally preferred to B demonstrating that the statistical distribution of linguistic structures within the corpus is adequate: the iterative analysis of the corpus produces similar results to the global learning strategy offered by SNNs.

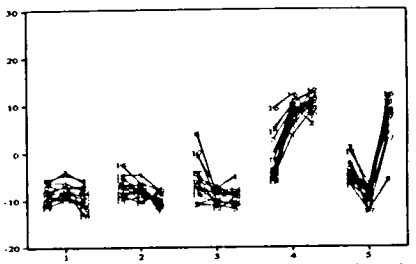


Figure 3. Superposition of syllabic f0 contours for 17 incredulous questions.

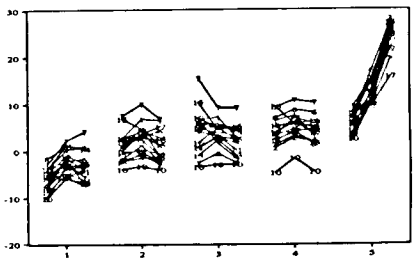


Figure 4. Superposition of syllabic f0 contours for the same set of sentences as Fig.3 but uttered as full questions.

FURTHER WORK

A major challenge to our approach is to demonstrate that prosodic information is distributed along the phonetic string via global contours which could be extracted and compared to melodic prototypes. We try here, to extend the proposal made by Fónagy [6] using the term "clichés mélodiques" for large pieces of prototypic melodic shapes associated to given communication needs.

We are thus looking for prosodic events associated with sentence level that can not be explained by the linguistic sub-unit configurations, e.g. an utterance template more elaborate than the two basic declination lines currently used. We thus recorded a second corpus designed especially for revealing the existence of global prototypic contours associated with sentence level: this corpus consists of short utterances (from 4 to 6 syllables) in order to limit the influence of carried sub-units. The unmarked syntax allowed several modalities such as assertion, question, exclamation... with various affects such as incredulity, doubt, surprise...

Early results with 5 syllables sentences seem to confirm our predictions. Figures 3 and 4 respectively show the superposition of 17 incredulous vs full questions ("Question incrédule" mentioned by Fónagy [6]) with various syntactic and phonotactic structures[1]. Note in Figure 3 the emergence of a global melodic prototype at the sentence level with an "accent" on the penultimate syllable.

CONCLUSION , PERSPECTIVES

Preliminary results show that both generation methods presented in this article are well appropriated for automatic generation of F0 contours using our theoretical framework.

The structured lexicon enables the main differences between prosodic movements of the same phonological level to be captured and allows one to observe the way in which prosodic contours of different phonological levels are combined.

The SNN approach seems the most promising, since it achieves a high quality of synthetic F0 curves with fewer assumptions on the shapes of elementary patterns. Moreover, this approach is a very versatile one: such a strategy has been efficiently applied to rhythmic control [4] and will enable coherent multiparametric prosodic generation. The next step will be the training of a new sequential neural network with both rhythmic and melodic information patterns.

The automatic learning capacities of SNNs have to be guided by a strong hypothesis for the way linguistic units and affect are encoded onto the prosodic signals. Our assumptions is that this encoding is done via global patterns which could be quasi directly perceived and identified. This "direct perception" of intonation has immediate applications in the field of speech recognition.

[1] The number of words goes from 1 to 5. The relative size of each word was also varied within each syntactic structure.

REFERENCES

[1] Aubergé, V. (1992), Developing a structured lexicon for synthesis of prosody. In Bailly, G.Benoît, C., editors, *Talking Machines: Theories, Models and Design*, pp. 307-321. Elsevier B.V.

[2] Bailly, G. (1989), Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, vol. 8, pp. 137-146.

[3] Bailly, G. (1992), Barbe, T. and Wang, H. Automatic labelling of large prosodic databases: tools, methodology and links with text-to-speech system. In Bailly, G.Benoît, C., editors, *Talking Machines: Theories, Models and Design*, pp. 323-333. Elsevier B.V.

[4] Barbosa, P. and Bailly, G. (1994), Characterisation of rhythmic patterns for text-to-speech synthesis. *Speech Communication*, vol. 15, pp. 127-137.

[5] Emerard, F. and Benoît, C. (1987), De la production à l'extraction, l'état d'un chantier, *16èmes journées d'Etudes sur la Parole*, pp. 224-228.

[6] Fónagy, I., Bérard, E. and Fónagy, J. (1989), Clichés mélodiques. *Folia Linguistica*, vol. 17, pp. 153-185.

[7] Grosjean, F. (1983), How long is the sentence ? Prediction and prosody in the on-line processing of language, *Linguistica*, vol. 21, pp. 501-529.

[8] Jordan, M.I. (1989), Serial order: A parallel, distributed processing approach. In Elman, J.L. and Rumelhart, D.E., editors, *Advances in connectionist Theory: Speech*. Lawrence Erlbaum, Hillsdale, NJ.

[9] Ladd, D.R. (1983), Phonological features of intonation peaks. *Language*, vol. 59(4), pp. 721-759.

[10] Silverman, K.E.A. & al (1992), TOBI: A standard for labelling English prosody, *Proc. International Conference on Spoken Language Processing*, vol. 2, pp. 867-870.

[11] Traber, C. (1992), F0 generation with a database of natural F0 patterns and with neural network. In Bailly, G.Benoît, C., editors, *Talking Machines: Theories, Models and Design*, pp. 287-304. Elsevier B.V.

PITCH CONTOUR STYLIZATION USING A TONAL PERCEPTION MODEL

P. Mertens and Ch. d'Alessandro
CCL, University of Leuven, Belgium and LIMSI, Orsay, France

ABSTRACT

An algorithm for automatic pitch contour stylization is described. It is based on a model of tonal perception, such that the resulting stylization is controlled by two perceptual thresholds. The output is the sequence of audible pitch events aligned with the syllables in the utterance.

INTRODUCTION

Stylization is a manual or automatic procedure that modifies the measured F0 contour of an utterance into a simplified but functionally equivalent form, i.e. preserving all melodic information which has a function in speech communication. There are several motivations for doing this; for instance, to reduce the amount of data required to generate pitch contours in synthetic speech, or to isolate the functional parts in the contour (and to remove the others) and to obtain a representation of this underlying contour, e.g. for intonation teaching or linguistic research. We propose an approach in which the stylization represents the pitch contour which is perceived by the average listener. In other words, the stylization process is seen as a simulation of tonal perception and as a way to measure what is heard. This approach satisfies both goals mentioned earlier: it results in an important data reduction and it filters out F0 events which cannot be heard and hence have no function in prosody. Still, there are other motivations for computing the perceived pitch: as a way to evaluate phonological intonation models and to obtain an automatic transcription of intonation. This may require some explanation.

There is little doubt about the acoustic manifestation of intonation (at least, for pitch): for most speech signals, F0 can be computed in an objective way, with estimation errors below perceptual thresholds. The phonological representation, however, is a sequence of symbols (tones, pitch movements) the determination of which involves a phonetician who interprets the data within a particular model. The large number of such models suggests the lack of a procedure to evaluate them. How could one decide that the descriptive units of model A are more viable than those of model B, or even that they are psychologically viable at all ? To this date, there is no clear criterion for the verification of intonation models; as a result their choice is often a matter of personal preference.

When someone describes the sounds he hears, he refers to the auditory image resulting from sensory and perceptual processing, rather than to the acoustic signal. It can easily be seen that the cognitive process of intonation understanding does not have direct access to the acoustic form (F0) but rather to the pitch events after processing by the peripheral auditory system and the perceptual system. Consequently it is this form that should be the input to the phonological model. By computing this perceived pitch contour, one will narrow the gap between the acoustic and the cognitive domain, because it eliminates one of the assumptions made by phonological models (namely the one about the nature of the input representation).

The rest of this paper gives a quick overview of tonal perception effects, then describes the algorithm implementing them. Finally we describe some of the results obtained.

Tonal perception

What is known about tonal perception?

1. Spectral and amplitude variations in the speech signal affect the way in which pitch variations are perceived, giving rise to a perceptual segmentation of the speech continuum [5]. This *segmentation effect* results in a sequence of short tonal events aligned with the syllables, rather than a continuous pitch curve for the whole utterance. Unfortunately, no quantitative model describing the contribution of changes in (global) amplitude, spectral energy, and other attributes, is yet available.

2. The perception of a changing pitch requires some minimal amount of frequency change as a function of time. Otherwise a static pitch is perceived. This effect is known as the *glissando threshold* (*G*). For a uniform pitch change (with constant slope), $G = 0.16 / T^2$, where T is the duration of the pitch variation. The effect has been investigated for years [3,4,8,9], both for pure tones and synthetic speech, but not in continuous speech.

3. A change in pitch slope will be perceived provided some minimal difference in slope, known as the *differential glissando threshold* (*DG*). There has been little research on this effect [4].

4. Static tones, i.e. short-term F0 variations which are below threshold G, are perceived with a certain pitch. In a study on the perception of vibrato [1], it was shown that this *short-term integration* can be modelled by a windowed time average (WTA) function.

Our stylization algorithm simulates these four effects.

DESCRIPTION OF THE STYLIZATION ALGORITHM

Figure 1 shows a block diagram of the stylization algorithm. It consists of several processing steps, some of which are purely acoustic (F0 measurement, voicing determination), while others are related to tonal perception. We will focus on the latter here. Most of the work in the algorithm concerns the determination of the speech fragments for which the perceptual effects are to be computed.

1. *Speech segmentation.* Since pitch perception is determined by spectral and amplitude changes, the speech signal is first divided into syllable-sized chunks. In the absence of a quantitative model of this effect, several types of segmentation are investigated. The first focusses on spectral change and uses the voiced parts of the syllables [2]; the second favours amplitude change and computes the syllabic nuclei (or loudness peaks) [6]. We will illustrate the results obtained with both segmentations.

2. *Short-term perceptual integration* of pitch. The WTA model is applied to the F0 in the voiced region of each syllable. This results in a smoothed pitch contour (as can be seen in Figure 2).

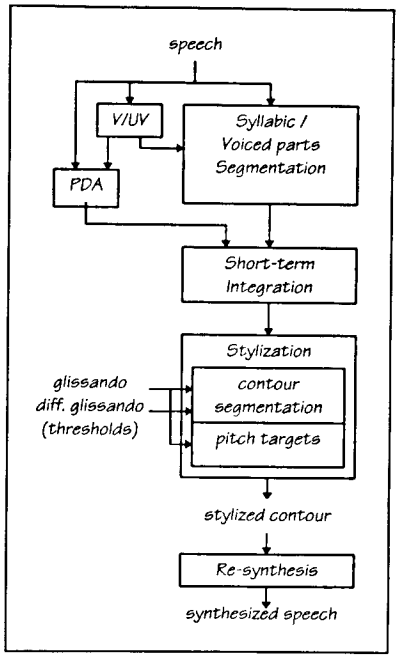


Figure 1. Block diagram of the pitch contour stylization algorithm. PDA is the pitch determination algorithm. V/UV is the voicing decision.

3. *Syllabic contour segmentation.* Syllabic pitch contours can be compound (e.g. rise-fall); they should be divided into simple, uniform parts first. This results in one or more *tonal segments* per syllable: a monotonous pitch change, i.e. either level, rising, or falling, and without an audible change in slope. This segmentation is motivated by the fact that the *G* and the *DG* are obtained for, and should therefore be applied to such uniform segments.

The syllabic contour segmentation involves two steps. The first locates the turning points in the contour so as to break it down into candidate tonal segments. The second makes a decision as to which candidate segments are to be grouped. The first step is recursive. Within an analysis interval with an audible pitch change (above *G*), a new turning point is found at the point of

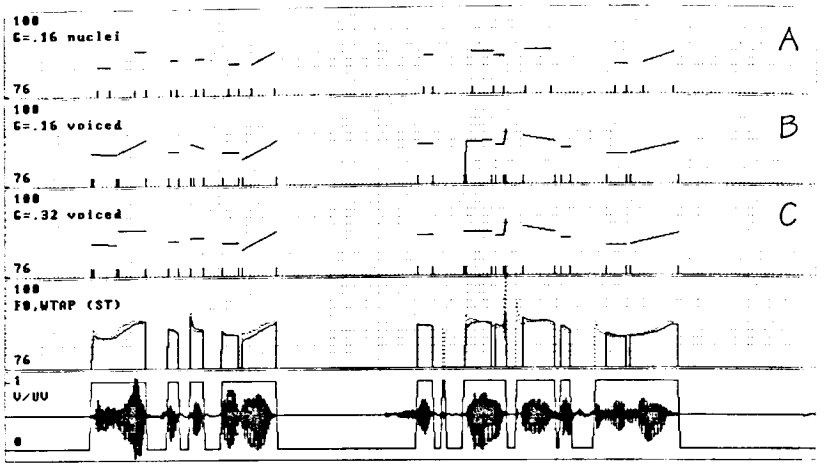


Figure 2. Comparison of stylization results with different parameter settings, for the utterance "d"ailleurs quand t'es pas là, [pause] tu sais de quoi on cause entre nanas", with a signal duration of 3.58s.. See text for explanation.

maximum difference between the observed (WTA-)pitch contour and a straight line between the start and end pitch, provided this difference exceeds some critical value of 1 ST. When a turning point is found the same procedure is applied recursively to both parts, before and after the turning point. The second step merges two consecutive candidate segments if their slope difference is smaller than the *DG*. It proceeds from left to right. After each merger the list of candidate segments is updated.

An interesting property of this procedure is that it is entirely controlled by two parameters, which are perceptual thresholds *G* and *DG*.

4. *Stylization*. For each tonal segment obtained in the previous step, the estimated *pitch targets* are obtained as the (WTA-)pitch at the start and end of the tonal segment. The stylized contour is the linear interpolation between successive pitch targets. For static tonal segments the pitch of the end point is extended to the entire tonal segment.

RESULTS

In order to evaluate the stylization, the speech signal was resynthesized (TD-PSOLA) with the stylized pitch contour and presented to 20 listeners for comparison with the original signal. The results of this experiment are described in [2]. By changing the two parameters of the model (*G* and *DG*) it is possible to evaluate their impact on the stylization. By a systematical evaluation of the parameter settings in listening tests the system can be used to determine the thresholds *G* and *DG* in continuous speech.

Figure 2 shows the stylizations obtained with different types of segmentation and different settings of the model parameters *G* and *DG*. The figure contains five parts. The lower part displays the speech signal together with the V/UV decision. All others parts use a semi-tone (ST) scale for the Y axis, with grid lines 2 ST apart. The next part shows the F0 (dotted line) and the WTA pitch (full line). The latter is calculated on the F0 values in the voiced part of the syllable. This results in a smoothed and somewhat time delayed version of the F0. The three upper parts show stylizations obtained with different parameter settings. The upper stylization (**A**) uses a segmentation into syllabic nuclei and the "standard" glissando threshold *G*=.16 (this value is the numerator in the equation given earlier).

The small vertical marks delimit the tonal segments. The two other stylizations (**B**,**C**) use the segmentation into voiced syllabic parts with parameter *G*=.16 or *G*=.32, respectively. By using *G*=.32 the glissando threshold is doubled, simulating the hypothesis that in continuous speech pitch changes twice as large as those for isolated stimuli would be required in order to be audible. All three stylizations are obtained with parameter *DG*=20, although a setting of *DG*=40 produces the same result. Stylization **B** gives the largest number of dynamic tones, while in **C**, due to the higher value of *G*, two dynamic syllables have been stylized as static ones. In **A** there are even less dynamic syllables because the nuclei are generally shorter than the voiced parts (which has an impact on the *G*). The second part of the utterance contains a F0 detection error, which is present as a dynamic tonal segment in **B**, and as a static one in **C**.

DISCUSSION

The stylization based on tonal perception has several inherent advantages over other types of stylizations.

1. It gives both a *qualitative* and a *quantitative representation* of the auditory contour, showing *how* the contour is perceived (which parts are perceived as dynamic, which as static, and which parts are not audible), and *what pitch* is perceived, for any time instant *t*. While many stylizations are descriptively adequate, ours also offers explanatory adequacy. In this respect, pitch movement approaches (e.g. "close-copy stylization") are less elegant because they sometimes suggest that the listener hears a changing pitch (in unvoiced syllable onsets, e.g.) where actually he hears no pitch at all.

2. It is *theory-independent*: it isn't linked to a particular prosodic model; it doesn't refer to pitch levels (which would have to be identified), to a declination line (which would have to be determined), to normalized pitch movements or contours, and so on.

3. The stylization proceeds from *left to right*, and can be applied to speech fragments as small as syllables. As a result one does not need the entire utterance to calculate the stylization (as is the case for declination line based procedures).

The stylization can be used as a tool for basic research.

1. By varying the model parameters, in combination with resynthesis and listening tasks, it can be used to measure the perceptual thresholds *G* and *DG* for continuous speech, at least for speech signals with an obvious segmentation into syllables.

2. The stylization provides an automatic transcription of the perceived tonal events, while eliminating the bias of the human transcriber. As such it is useful in the construction and verification of prosodic models.

REFERENCES

[1] Alessandro, C. d' & Castellengo, M. (1994), "The pitch of short-duraton vibrato tones", *J. Acoust. Soc. Am.* 95, pp. 1617-1630.

[2] Alessandro, C. d'; Mertens, P. (forthcoming, 1995), "Automatic pitch contour stylization using a model of tonal perception", *Computer Speech and Language*.

[3] Hart, J. 't (1976), "Psychoacoustic backgrounds of pitch contour stylisation", *I.P.O.- Annual Progress Report* 11, pp. 11-19.

[4] Hart, J. 't; Collier, R. & Cohen, A. (1990), *A perceptual study of intonation*, Cambridge: Cambridge Univ. Press, 227 pp.

[5] House, D. (1990), *Tonal Perception in Speech*, Lund: Lund Univ. Press

[6] Mertens, P. (1987), "Automatic segmentation of speech into syllables", in Laver, J. & Jack, M.A. (eds.), *Proc. of the European Conf. on Speech Technology*, vol. II, 9-12.

[7] Mertens, P. (1989), "Automatic recognition of intonation in French and Dutch", *Proc. Eurospeech* 89, vol 1, pp. 46-50.

[8] Rossi, M. (1971), "Le seuil de glissando ou seuil de perception des variations tonales pour la parole", *Phonetica* 23, pp. 1-33.

[9] Rossi, M. (1978), "La perception des glissandos descendants dans les contours prosodiques", *Phonetica* 35, pp. 11-40.

ON THE ANALYSIS OF SYLLABLE TIMING IN EVERYDAY SPEECH

Henrietta J. Cedergren and Hélène Perreault
Département de Linguistique
Université du Québec à Montréal
C.P. 8888, Montréal H3C 3P8, Québec, Canada

ABSTRACT

A regression analysis is presented which investigates the effects of surface prosodic structure features and speech tempo on syllable timing in Montreal French. A cluster analysis of the 16 speakers' regression coefficient estimates allows us to distinguish between patterns of effects that are systematic across speakers, and patterns of effects that are speaker specific.

INTRODUCTION

Explicating temporal organization in speech is a complex task which involves extricating the effects of multiple parameters. Previous experimental research has shown that observed patterns of timing may follow from properties of the linguistic text, i.e. segmental and prosodic organization features [1, 2], properties of the informational content, i.e. the flow of information in discourse [3, 4], or properties of the context, i.e. citation vs spontaneous speech forms [5]. Experimental paradigms have thus allowed researchers to investigate specific properties while presumably controlling for other dimensions of variability and have provided insights on the relevant parameters involved in the prediction of timing. However the relation between models of timing in experimental contexts and the timing of natural speech remains to be clarified. Little is known, for example, about the relations between observed properties of timing and inter-speaker variation characteristics, i.e. socio-symbolic differences among speakers [6].

This study is concerned with addressing the issue of understanding the relation between temporal organisation and inter-speaker differences in speaking style in everyday speech. It bears on the problem of modelling timing in everyday speech. It attempts to distinguish between patterns of effects that are systematic across speakers, and patterns of effects that are speaker specific.

METHODS

The database for this study consists of three minutes excerpts of running speech extracted from hour-long recordings of sociolinguistics interviews of Montreal French. A sample of sixteen speakers differentiated according to sex, age and social class was used in the analysis. Speaker ages correspond to two generational categories: twenty to twenty five years of age or fifty five and over; an equal number of working class and middle class speakers were selected.

Table 1. Distribution of speakers according to sex, age and social class.

	working class		middle class	
sex	M	W	M	W
young	2 23	7 50	25 113	43 70
old	28 37	45 107	75 81	31 61

The speech excerpts were sampled at a rate of 16kHz/s. Segment durations were measured by manually placed cursors on a spectrogram time-aligned with a waveform; segment durations and labels were stored in an automatically generated file which was also coded for perceived prominence and prosodic grouping. Three levels of prosodic organisation were identified: phonetic syllables, rhythmic groups and intonational phrases. Prominence was distinguished as either demarcative, associated with the right boundary syllable of rhythmic groups or intonational phrases, or non-demarcative, secondary prominence, associated with non-final syllables.

A duration model

As reported elsewhere [7] we have been concerned with modelling the relation between observed syllable durations in milliseconds and features of surface prosodic structure and syllable composition. In the present analysis, the factorial set codes for eight infrasyllabic, suprasyllabic and tempo effects. At the level of the syllable, onset and rhyme complexity are coded in terms of the number of segments. Suprasyllabic prosodic features distinguish positional effects within the intonational phrase and the rhythm group, and prominence. Tempo is coded as a local measure of articulation rate within the intonational phrase [7]. Finally, we include a multiplicative factor which accounts for the interaction of infrasyllabic complexity in the rhyme with intonational phrase final position.

Table 2. Factorial categories.

Onset complexity in number of segments	OC
Rhyme complexity in number of segments	RC
Intonational phrase penult position	IPP
Intonational phrase final position	IPF
Non-demarcative prominence	NDP
Rhythmic group final position	RGF
Intonational phrase tempo	IPT

Interspeaker differences and similarities

We address the issue of modelling interspeaker differences and similarities based on the SAS complete linkage cluster analysis method. The regression coefficient estimates derived for each speaker's syllable duration model were Z-score normalized. These served as input to the cluster procedure.

RESULTS

The linear regression analyses revealed differences in temporal effects of the factorial scheme previously described. Speakers in the sample differ both in how the factors jointly account for observed syllable duration in their speech ranging from 48.36% for Spkr 25 to 67.43% for Spkr 75, and in how the proposed effects achieve statistical significance ranging from a maximum of eight factors to a minimum of 5 core factors. Thus, OC, RC and IPT emerged as significant predictors of syllable timing for all speakers. While IPF and IPF*RC assume different speaker specific patterns. Both predictors are significant for 6 speakers. IPF*RC was significant for 8 speakers, while IPF was not. The reverse pattern was obtained for 1 speaker and finally neither were significant for 1 speaker.

The following equation illustrates how the factors account for 57.38% of the variance of one of the speakers, Spkr 2, a young working class male:

Syllable duration =
constant + 51.67OC + 46.36RC + 99.05IPT + 27.56IPF + 8.20RC*IPF + 21.06IPP + 12.35NDP + 7.49RGF

Cluster analysis further investigated the issue of interspeaker temporal differences and similarities using as input data each speaker's coefficient estimates. The results of the complete linkage clustering procedure [8] revealed that the sample of sixteen speakers could be

divided into four groups based on their linguistic similarities: group 1 consisting of 6 speakers (25, 61, 50, 113, 2, 7), group 2 consisting of 5 speakers (23, 28, 45, 37 70), group 3 consisting of 4 speakers (31, 75, 81, 107) and finally group 4 restricted to 1 speaker (43). The hierarchical grouping of speakers that was obtained by measuring the maximum Euclidean distance among the clusters is illustrated in Figure 1. A comparison between the grouping illustrated in Figure 1 and the pre-analytic grouping of speakers displayed in Table 1 suggests that syllable timing behaviour is not a straightforward mirror of socio-demographic grouping. However certain observation can be made about the predominant socio-demographic characteristics of the first three clusters. Two of the three groups appear to be predominantly defined by age. Five of the six speakers in group 1 are young; all of the four speakers in group 3 are older. Social class appears to be associated with speaker differentiation in groups 2 and 3. Four of the five speakers in group 2 are working class; three of the four speakers in group 3 are middle class.

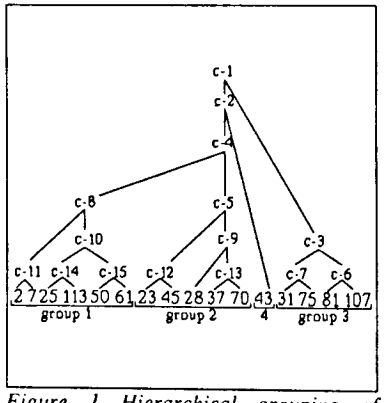


Figure 1 Hierarchical grouping of speakers.

Although the clustering procedure indicates that speakers can be allocated to four groups based on their temporal behaviour, it falls short in addressing further questions such as: How do the particular prosodic and tempo effects contribute to characterize each group? We attempted to explore this issue by means of the least squares comparison of means option of the GLM procedure in SAS and thus evaluate the null hypothesis that the difference in means among the standardised coefficient values for each factor for each pairwise comparison of the four clusters is not significant. Table 3 summarizes the results of the tests.

It appears that intergroup discrimination is differentially associated with the prosodic and tempo parameters. At one level, these tests reveal that certain temporal characteristics are shared by all speakers in the sample and could be construed as a shared dialect feature. Thus the null hypothesis is not rejected in all pairwise comparisons of the OC, RGF and RC*IPF effects. At the opposite extreme, these tests reveal that the most consistent indicator of group membership appears to be the local measure of tempo effect (IPT) which significantly distinguishes all pairwise comparisons of groups of speakers except groups 2 and 4. The other parameters have intermediate discriminatory effects. Some appear to be group specific; the IPP effect discriminates group 4 from all other groups of speakers. Others are more limited in the specificity of their discrimination among the groups of speakers. Thus group 1 is distinguished from groups 3 and 2 by NPD; IPF distinguishes group 2 from groups 1 and 4, and RC distinguishes group 3 from groups 1 and 2.

DISCUSSION

We have presented some preliminary results of an investigation which aims to understand the relation between temporal organization in everyday speech and interspeaker differences in speaking style. A factorial scheme of eight prosodic and

Table 3. Results of least squares means comparisons of speaker groups by prosodic parameters. (= significant comparison)*

	OC	RC	IPP	IPF	NDP	RGF	IPT	IPF*RC
1 vs 2					*		*	
1 vs 3		*			*		*	
1 vs 4			*				*	
2 vs 3		*		*			*	
2 vs 4			*	*				
3 vs 4			*				*	

tempo parameters was found to account for the observed variance in measured syllable duration ranging from 48.36% to 67.43%.

The use of cluster analysis has allowed us to examine the issue of how speakers are grouped based on their timing behaviour. A post-hoc inspection of the socio-demographic characteristics of the speakers in each cluster group revealed that both age and social class appear to be related to the grouping. These results suggest that differences in syllable timing in spontaneous speech are determined not only by linguistic properties, but that they may also reflect inter-speaker socio-demographic differences.

Exploratory analysis of each cluster using a least squares means comparison revealed that prosodic and tempo parameters do not operate uniformly in discriminating among groups of speakers.

Although these results are based on the analysis of a small sample of sixteen speakers, we are confident that they provide substantive evidence that temporal organization is an important component of speaking style.

ACKNOWLEDGEMENT

This research was supported by the Social Sciences and Humanities Research Council of Canada under Grant No. 410-92-1840.

REFERENCES

[1] Lehiste, I. (1975), The phonetic structure of paragraphs, *Structure and Process in Speech Perception*, ed. by A. Cohen & S. Nooteboom, Springer, Berlin. pp. 195-206.

[2] Campbell, W.N. (1992), *Multi-level Timing in Speech*, Unpublished PhD Thesis, Sussex University, Department of Experimental Psychology.

[3] Bruce, G. & P. Touati (1992), On the analysis of prosody in spontaneous speech with exemplifications from Swedish and French, *Speech Communication* 11, pp. 453-458.

[4] Van Santen, J.P.H. (1992), Contextual Effects on Vowel Duration, *Speech Communication*, February 1992.

[5] Lindblom, B. (1990), Explaining phonetic variation: A sketch of the H & H theory, in *Speech Production and Speech Modelling*, ed. by W. Hardcastle and A. Marchal, Kluwer Academic Publishers, Dordrecht, pp. 403-439.

[6] Labov, W. (1994), *Principles of linguistic change: internal factors.* Blackwell, Oxford.

[7] Cedergren, H. J. & H. Perreault (1994), Speech Rate and Syllable Timing in Spontaneous Speech, *Proceedings of the International Conference on Spoken Language Processing*, vol.3, pp. 1087-90.

[8] Anderberg, M. R. (1973), *Cluster Analysis for Applications*, New York: Academic Press inc.

WHAT DO SPECTRAL AND PERCEPTUAL ANALYSES REVEAL ABOUT SPONTANEOUS SPEECH IN DIALOGUES OF DIFFERENT STYLE?

Lioba Faust
Institut für Kommunikationsforschung und Phonetik
Poppelsdorfer Allee 47, 53115 Bonn, Germany
Phone: +49-228-735641 E-Mail: lfa@asl1.ikp.uni-bonn.de

ABSTRACT

The analyses of segmental changes, vowel durations of [i:, ɪ] and [a, ɑ:], and listeners' classifications of utterances of three different variants of German dialogues reveal that there are no crucial characteristics to classify speech as either spontaneous or read. The casual spontaneous style shows the strongest phoneme and syllable reductions and is generally classified correctly. For the careful spontaneous and for the read utterances listeners' classifications vary strongly with respect to the speakers.

INTRODUCTION

In automatic speech recognition a shift of interest can be observed from read speech towards spontaneous speech. The variability of the speech signal is expected to be higher in spontaneous than in read speech and even higher in casual than in careful spontaneous speech. Nevertheless, linear modifications between the styles cannot be assumed [2]. Furthermore, read speech must not be considered as a contrast to spontaneous speech, but may show as much stylistic variation as spontaneous speech [3]. Therefore, the purpose of the experiment was to have a look at variability in three *natural types of conversation* together with as little restrictions as possible on the controlled elicitation of speech, which is both of great phonetical interest and indispensable for an automatic treatment of everyday conversation. For the same reason, we accepted the fact that speech material would be linguistically and phonetically different and rejected the restriction on relatively small units of speech, e.g. sequences of sounds or isolated sentences.

METHOD OF EXPERIMENTATION

Corpus design and recordings

Four female German students between 23 and 27 years of age participated in an experiment of dialogue recordings of the following different speaking styles: 1. casual speech: totally free conversation, 2. careful spontaneous speech: time-scheduling negotiation dialogue using the formal mode of address with the relevant dates given in a calendar, 3. read speech: rereading of the transcribed utterances of the second dialogue variant. Hesitations, word repetitions and repairs that had been transcribed were generally dropped for the copy to be reread. The most important issue for the rereading copy was to preserve the dialogue structure and, in general, the grammatical structure of the utterances.

The speakers were sitting in the institute's speech laboratory, in two neighbouring rooms separated by a glass pane. The communication was performed using headset microphones (Sennheiser HMD 414-6). Speech was digitally recorded on separate channels.

The casual dialogues were recorded first without the speakers' knowledge of being recorded. While the supervisor went away under a pretext for about three minutes, the speakers where left on their own. This led to very different dialogue structures. Moreover, the acoustic conditions such as a constant distance between the speakers' mouth and the microphone could no longer be controlled. This was accepted, since speech under close to natural conditions was desired. For the same reason, we could not use the headset, therefore two condenser microphones (Neumann KM 140) were used. For spectral analyses as well as for the listeners' experiment acoustically "useful" utterances had to be very carefully selected.

Listening experiment: Design and performance

From each speaker and each variant three utterances of phrasal length were selected. The utterances were grammatically sentence-like, with different intonational patterns, and they contained a certain amount of pauses, hesitations and repairs. From the careful spontaneous and the read versions, identical utterances were collected. The utterances were presented in random order, and each utterance was played twice.

The listeners had to classify the utterances as "spontaneous" or "read" in a forced choice task. Furthermore, the listeners were asked to rate the degree of reliability for their decision on a five-point scale from very safe to very unsafe. Moreover, the essential linguistic or phonetic features underlying the listeners' decisions had to be specified. The given criteria were: syntactical structure, speech fluency, repairs, articulation, intonation, and speech rate.

The classifications were made by ten phonetically educated listeners who were members of the institute and twelve naive listeners (beginning students).

Spectral and segmental analyses

All dialogue utterances were orthographically transcribed, manually segmented and labelled and marked with phrasal accents. The spectral analyses that were necessary for the examination of segment durations and phoneme productions were carried out using a PC programme for speech labelling developed at the IKP (SONA) [5].

Vowel duration

The duration of the vowels [i:, ɪ] and [ɑ:, a] was measured, and the mean value was calculated for each speaker and each speaking style using SPSS for PC. First, two groups were built for each vowel: phrase-final and non phrase-final. Then, each group was subdivided into phrase-accented and non-accented vowels. Finally, the groups for non-accented vowels were divided into function words and content words. This grouping was made with regard to the prosodic variation that was observed for the different speaking styles.

Segmental reduction

Reduction of segments was measured by comparing the string of labelled speech with the canonical phoneme symbols that were obtained from the orthographic transcription using the PC programme P-TRA developed at the IKP [6]. Segmental variations were grouped into deletion of syllables (including contraction of words), deletion of sounds, substitutions and insertions. The number of misarticulations was calculated for each speaker and each speaking style (SPSS).

RESULTS

Listening experiment

Table 1: Number of "spontaneous" and "read" classifications per speaker and style (Sp. = Speaker; cas. = casual; car. = careful; r = read)

	Utterances classified as spontaneous			read			
Sp.	cas.	car.	r.	cas.	car.	r.	Total
BP	37	52	30	06	14	35	174
SO	64	62	25	02	04	41	198
VB	58	41	37	06	25	29	194
VG	59	58	27	06	08	39	197
Total	218	213	119	20	51	144	763

As illustrated in Table 1, the casual utterances were correctly identified in almost all cases. Wrong decisions were in all cases reached by naive listeners. The appearing clear correct decision for this style may also be seen as a fact of the restricted acoustic conditions that had been present and that was certainly more easily perceivable by the educated listeners.

Having a look at the results for careful and read speech, correct decisions seem to be predominant. A closer look at the

individual speakers concerning the careful style shows a high degree of misclassifications for speaker VB, whereas for SO only 4 wrong decisions - reached by naive listeners - occur. For VB, both in careful and in read speech, educated listeners come to even more wrong decisions than naive listeners. For all speech styles, the right decision is significant (p=.0011). However, significance is no longer maintained by the educated listeners (p=.0957).

As to the reliability of listeners' decisions, listeners are generally very or rather safe about reaching a correct decision. This result is significant (p<.01), but depends on the speakers and the spealing style. Looking at the read utterances by speaker VB, educated listeners are in most cases undecided, regardless whether their decision is right or wrong, whereas naive listeners are in most cases rather safe even when their decision is wrong and rather safe or undecided when they are right. The results are similar looking at the careful utterances of the same speaker.

The obvious supposition is that educated listeners are more careful in classifying perceived utterances as "read" or "spontaneous" as they are more used to carefully listening and more conscious of the variability in speech utterances than naive listeners.

For the correct decisions, the distribution of the phonetic or linguistic features was examined. Concerning the casual style, all features are mentioned more or less frequently by the listeners with only "speech fluency" being significant (p= .0003). "Fluency" is also significantly often mentioned in careful style (p<.01). Other criteria here are "repairs" (p= .0021), "intonation" (p=.0140) and "articulation" (p<.01 for the group of educated listeners). For the read speech style "fluency", "articulation", "intonation", and "syntactical structure" yielded significant results (p<.05).

Vowel duration

Results on vowel duration show a great variability depending on word category and phrasal accentuation and are of course very speaker dependent. For the individual speaker a clear trend concerning all vowel groups cannot be observed. This means that a single style is to be seen as a unique expression of speech, and vowel duration alone cannot generally indicate a specific speech style. The results show that casual speech is not obviously faster, and read speech, as it might be most carefully articulated, is not necessarily slower than careful spontaneous speech.

For [ɪ], results (cf. Table 2) support the classifications of the listening experiment, i.e. speaker SO slows down as she speaks more carefully whereas for VB mean durations become shorter. The results of BP and VG correspond to the listeners' decisions as duration in most cases does not change very strongly. Moreover, the large standard deviation in all cases indicates that the variation of speech rate differs to a great extent within a single phrase. Results are similar for the other vowel groups that had been analysed. They are not listed for reasons of space.

Table 2: Mean and standard deviation (ms) of vowel durations for non-final, non-accented [ɪ] in function words

	casual	careful	read
Sp.	Mean; std	Mean; std	Mean; std
BP	56,7; 23,1	60,1; 16,1	53,6; 16,3
SO	55,9; 24,6	56,7; 21,5	65,8; 19,4
VB	60,3; 21,9	48,5; 17,0	54,6; 16,8
VG	58,1; 24,1	55,7; 18,1	47,2; 12,5

Segmental reduction

For all speakers, the greatest difference in articulation compared to the canonical form is found in casual style. Here we find a large amount of syllable deletions, but also sound deletions (in most cases final stops) and substitutions. The only sound changes that are found less frequently than in careful and read style are insertions.

The difference between careful and read style is exemplified in Figure 1. For speaker SO an impressive decrease of deletion can be noted, for VB instead, an increase towards read speech. For the speakers BP and VG sound changes are not so heavy. These results explain the listeners' decisions in the listening experiment.

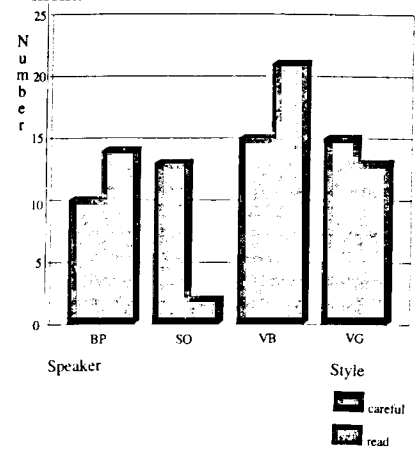


Figure 1: Sum of syllable deletions

CONCLUSION AND DISCUSSION

Listeners' decisions, especially for the speakers SO and VB, show that there exists a clear concept of what is expected both for read and for spontaneous speech. Obviously, read speech is expected to be very clear, i.e. without any misarticulations or repairs, and fluent, which means absence of pauses and hesitations. The appearance of hypercorrect articulation is related to read speech whereas a certain amount of segmental reduction is associated with spontaneous speech. If this concept cannot be recognized clearly, utterances cannot be unambiguously classified.

Of course, apart from the selection of utterances and the acoustical conditions, reading skills may influence the results, especially concerning the listening experiment. If reading skills are to be defined as the ability of perfectly transferring visual print patterns into acoustic patterns, then the utterances of speaker SO should be considered as a perfect reading. Listening to VB, however, utterances are very fluent and sound more natural than those of SO.

As a conclusion, these results suggest that the way listeners expect read speech to be does not match reality for all speakers. Moreover, the degree of acceptance of a speaker's reading style might be low. Read speech should rather be considered as having as much stylistic variation as spontaneous speech. Casual speech, on the other hand, seems to be a kind of slurred variation of spontaneous speech. This means that any expression of spontaneous or read speech is adapted to the given particular communicative situation, which is performed speaker-specifically.

Further examination of speech rate within a phrase, dynamic range, range of F0 and listeners' classifications of manipulated utterances, e.g. inserting pauses and hesitations into the signal of read utterances is called for. Furthermore, the classification of utterances related to a specific style may yield interesting results.

ACKNOWLEDGEMENT

This research study was partly supported by the German Federal Ministry of Education, Science, Research and Technology (BMBF).

REFERENCES

[1] Blaauw, E. (1992): Phonetic differences between read and spontaneous speech. In: Proc. ICSLP, Vol. 1, Banff, 751-754.

[2] Eskénazi, M. (1992): Changing speech styles: Strategies in read speech and casual and careful spontaneous speech. In: Proc. ICSLP, Vol. 1, Banff, 755 - 758.

[3] Eskénazi, M. (1993): Trends in Speaking Styles Research. In: Proc. ESCA Eurospeech, Vol.1, 501-508. ESCA ETRW, Barcelona, 36.

[4] Swerts, M./Collier, R. (1992): On the controlled elicitation of spontaneous speech. In: Speech Communication 11, 463-468.

[5] Stock, D. (1994): SONA. http://asl1.ikp.uni-bonn.de//sona.html.

[6] Stock, D. (1992): P-TRA - Eine Programmiersprache zur phonetischen Transkription. In: Beiträge zur angewandten und experimentellen Phonetik. Steiner Stuttgart.

THE FORMS AND FUNCTIONS OF INTONATION IN THE PHONE VOICE

E. Douglas-Cowie and Roddy Cowie
School of English/Psychology, Queen's University Belfast

ABSTRACT

The paper describes patterns of pitch in secretarial business calls. Distinctive patterns are used in a formulaic way. Some are associated with particular parts of calls such as openings, transfers and closings; others mark transitions between key stages in calls.

INTRODUCTION

'The phone voice' is a commonplace expression. Presumably the expression reflects the fact that there is something distinctive about the way we speak on the phone. Much of what is distinctive undoubtedly lies at a prosodic level, and a few papers have begun to address prosody in phone conversations [1], [2], [3]. This paper adds to that body of knowledge. We focus on patterns of pitch in phone calls. These are known to have distinctive properties, but they remain partially described.

Several authors have noted the need to develop our understanding of intonation in interactive discourse [4], [5]. Phone calls are a useful context for that development because they offer prosodic patterns which are less easy to take for granted than most.

The study considers a specific class of calls, secretarial business calls. These presumably have special features of their own, but they also seem to encapsulate in extreme form what people describe as 'the phone voice'.

Our emphasis is on broad patterns within calls and their relation to communicative functions and roles. This is a necessary complement to close analysis of local features.

THE DATA

The study uses the speech of three secretaries in the School of Psychology at Queen's University Belfast. Each of the three recorded her own voice over a working day whenever she answered phonecalls or made them. Each secretary had control over the tape-recorder, so she could switch it off whenever she wanted e.g. during confidential calls. After some initial switching on and off, the secretaries tended to leave the tape-recorder running. The secretaries were chosen because auditorily they all appeared to have distinctive phone voices, though to varying degrees. A total of 82 calls were recorded. The voice of the other person on the line was not recorded, so the analysis is restricted to the secretaries' voices. The recordings also gave long samples of the secretaries' voices off the phone. These have been used as a reference point with which to compare the secretaries' phone voices.

RESULTS

Initial analysis revealed an overriding feature in the phonecalls - that they were highly formulaic. This was apparent in terms of both discourse structures and patterns of pitch, and there were clear relationships between formulaic aspects of the discourse and formulaic aspects of the pitch. These observations led to a second stage of analysis in which the observed structures of discourse and patterns of pitch were systematically examined. This was done by generating a template against which phone calls were systematically measured.

At a macro level calls were divided into a number of key functional stages. These are: (i) an opening phase characterised by formulaic greetings (e.g. 'School of Psychology' (ii) a post opening phase in which the function of the call is either signalled (in the case of initiated calls) or acknowledged (in answered calls) (iii) a transfer to another person (if required) (iv) an interaction stage for non transferred calls or for failed transfer calls (v) a preclosure phase in which a resolution or conclusion is reached (vi) a preclosure sequence of closing formulae (e.g. 'okay then, right, fine, okay') (vii) a final closing phrase.

Calls were divided into parts corresponding to these stages. For each part gross pitch descriptions were recorded. These noted overall pitch height, pitch variability and any pitch transition marking shifts to a new stage.

At a finer level, pitch patterns within each stage were more finely described in terms of shape, direction and patterning. Analysis was primarily auditory, but it was backed up by acoustic measurement when the need arose. Other relevant information was also recorded, in particular whether calls were initiated or answered, the main topic, the function of the call and the general mood.

The resulting records display clear relationships between formulaic aspects of the discourse and formulaic aspects of the pitch, at both macro and micro levels.

Macro structures

Macro levels are addressed first. The main point is that the majority of calls follow a broad formula in their structure, corresponding to the stages set out in the template. These stages appear to be marked by gross pitch signals involving overall pitch height and pitch transition.

The formula is not invariable. In particular, patterns of pitch height and pitch transition depend on whether the calls are answered or self-initiated. However both categories of call show a similar underlying pattern.

It is usually possible to identify a relatively constant central pitch within a stage: transitions between stages are marked by noticeable transitions in pitch height. These transitions themselves follow a larger pattern. They can either continue in the same pitch direction (upwards or downwards) for a series of stages, or they can show a type of yo yo effect where transitions move pitch between alternate levels.

Figure 1 plots pitch height against stage of call for a number of actual calls.

There are clear surface differences between answered and self initiated calls. Answered calls show a very regular pattern in their beginnings and endings. Pitch is characteristically high or very high in both. Self initiated calls end high to very high, like answered calls, but they tend to start start slightly lower.

In the portion between opening and closing, self initiated calls consistently show a concave pattern, dropping to the interaction stage and rising progressively through the three closing stages. The picture in answered calls is much less clear: some of the variation is considered below.

Several types of variation occur regularly.

The pattern of a clear central pitch within stages is linked to stereotypic interaction. When the secretary engages the listener to tackle a substantial task, pitch moves widely and unpredictably. This underlines the fact that stereotypy is a key issue in pitch patterning.

Some self initiated calls start quite low. Low openings occur in three particular types of calls: (i) continuations of earlier conversations between the secretary and the other person; (ii) calls where the secretary clearly expresses doubt or concern about how to resolve a central issue; (iii) calls where business is mixed with personal conversation.

A key variation in answered calls relates to a particular subtype, that is calls involving transfer to someone else. In these calls pitch is maintained high right from the opening stage to the point where the transfer is successfully executed (see calls Ans 1 and 2, marked with circles). This is unlike other answered calls where pitch drops sharply for the post opening phase (calls Ans 4 and 5).

Calls where transfer is attempted and fails form an interesting subgroup. Pitch is still maintained high right through any ensuing interaction (see call Ans 3). This continuation of the same pitch is as if the phonecall is held, as it were, at the opening stage: the secretary remains an intermediary whose role is to connect two other people, not a protagonist.

This suggestion fits evidence from calls which do not involve transfers, but follow a similar pitch structure. These involve situations where the secretary knows that she will be unable to deal with a query or to resolve the situation satisfactorily. Again, the secretary never fully engages in a satisfactory interaction.

Micro structures

Individual stages of phone calls are also formulaic, both in text and pitch patterns. We focus on openings, closings and transfers. These generally consist of stock, formulaic phrases. The patterns of pitch associated with them are similarly formulaic and distinctive.

Openings generally consist of stock greeting phrases and are marked by an

underlying pattern that juxtaposes fairly restricted pitch movement with extreme pitch movement. This holds in both answered and self intiated, with some variations. In answered calls, for example, the common pattern is a fairly level stretch followed by a very sharp rise to extremely high pitch (see Figure 2, left hand panel: the phrase is 'School of Psychology'). By comparison self initiated calls generally start with a level phase and end in a sharp fall.

Transfers also involve stock phrases accompanied by regular patterns of pitch. Some are addressed to the caller in phrases such as 'just hold on', and others to the object of the transfer (as in 'Peter, a call for you.'). Different patterns occur, but they are generally quite distinctive. They display sharp pitch movement, which may be preceded or followed by a phase of limited movement. Figure 2, right hand panel, shows an example for the phrase 'Just hold on a second please'.

Closings follow a distinctive formula involving two parts - first a build up of closing gestures (e.g. 'okay then, right, okay' etc) and then one final formula (usually in these calls, 'bye'). Figure 3 shows a typical closing pattern for the phrase 'That's fine, okay, bye.'

The closing gestures often show an alternating pattern where the pitch jumps up and down in quick succession. Pitch is also held much longer than normal on particular syllables, giving a singing-like quality. The final closing formula is usually marked by extreme pitch movement, classically on the word 'bye' where 'bye' is divided into two syllables, and the second syllable [e:] is held on a level pitch.

DISCUSSION

The use of pitch in phone calls appears to be distinctive, particularly at the level of formulaic patterns associated with particular stages. It is not clear whether similar patterns occur in other forms of speech: one might expect them to be extremes of a normal distribution.

One explanation for the patterns reflects a normal function of intonation, that is, to convey conversational signals e.g. opening, closing, taking turns etc. In the absence of visual cues, the load on these signals increases. Phone voice may provide enhanced intonational markers, and perhaps markers of types that are not normally be carried by intonation. Note also that cue reduction makes repair processes difficult on the phone. Conversational markers may be enhanced to avoid the need for repair. The general simplification and exaggeration of pitch patterns fit this account, and formulae may be seen as tried and tested solutions to a difficult problem.

A second theme is that formulaic patterns are a means of depersonalising interaction. They present prosodic features which reflect the speaker's role rather than her individual thoughts and character. The theme of depersonalisation provides a link between intonation and the segmental level, where the secretaries use forms which mask their local identity. For example, in normal conversation they all have marked forms of Ulster regional speech. But on the phone they use a much less regionally marked accent with partially Anglicised forms.

The element of role playing is not hypothetical. The secretaries told us that when they made calls they were not speaking as themselves, but as representatives of the organisation. In that capacity, they had two responsibilities. One was not to let their own personality interfere with the way they talked to callers. The second was to present a 'good' image of the organisation. We think this is an observation well worth following up.

REFERENCES

[1] Auer, P. (1988), "Rhythmic integration in phone closings", Working paper no. 2, Kontri, Univ. of Constance.
[2] Liberman, M. and McLemore, C. (1992), "Structure and intonation of business telephone openings", *Penn Rev. Linguistics*, vol. 16, pp. 68-83.
[3] Panese, M. (1994), "Prosody and Conversation. Phone-Closings in Radio-Talk", Working paper no. 24, Kontri, Univ. of Constance.
[4] Swerts, M. (1993), "Prosodic features of discourse units", Doctoral thesis, Technische Univ. Eindhoven.
[5] Couper-Kuhlen, E. and Selting, M. (1994), "Towards an interactional perspective on prosody and a prosodic perspective on interaction", Working paper no 29, Kontri, Univ. of Constance.

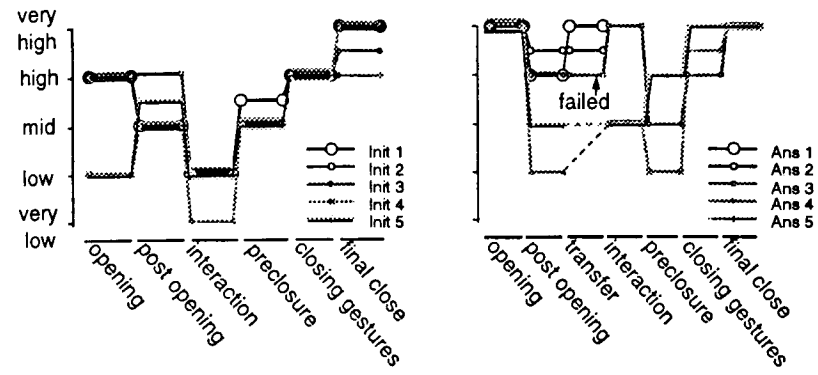


Figure 1: Pitch height and stage in representative calls, initiated by the subject (left hand panel) and answered by the subject (right hand panel).

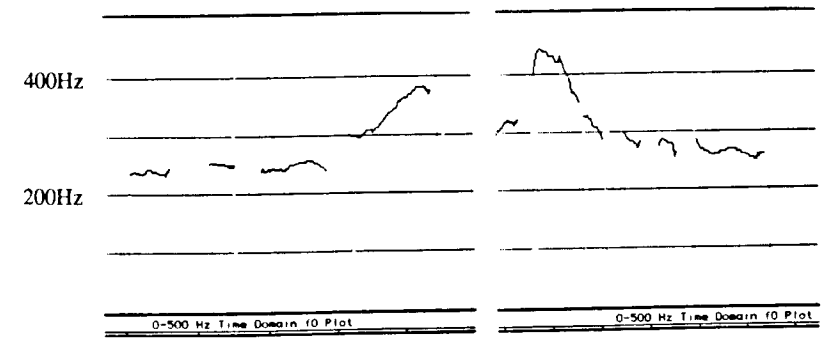


Figure 2: Opening greeting (left hand panel) and transfer (right hand panel)

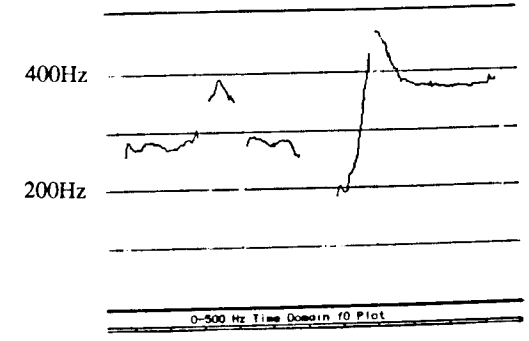


Figure 3: Closing gestures followed by final closing stage

PITCH RANGE AND REGISTER IN FRENCH POLITICAL SPEECH

Paul Touati
Dept. of Linguistics and Phonetics, Lund, Sweden

ABSTRACT

The aim of this paper is to present an analysis of pitch range and register of a French politician in two different political contexts. Methodological aspects will be first discussed. A contrast in pitch range and register along the hyper-hypospeech dimension is then proposed.

INTRODUCTION

Macrocontextual constraints such as general turn-taking conditions, dominance relationships between speakers, topic arrangements and rhetoric activity bear in a significant way on the use of a particular speaking style. French political speech is often characterized by a rhetoric use of acoustic-prosodic properties such as focal accent, contrasts in overall pitch and pauses (see [1] and [2]). Furthermore, prosodic correlates of a speaking style seems to be specified by the speaker in such a way that they are easily detected at a macrolevel by the listener (see [3]).

In [4], the issue of rhetorical prosody in political speaking style was addressed by analyzing contrasts in overall pitch as produced by a French politician (J. Chirac). As a result of this analysis, a two-fold categorization of overall pitch variation in French was proposed, one in terms of range and the other in terms of register. In the current study, overall pitch variation of another politician (R. Barre) has been analyzed along the time dimension represented by two different contexts a pre-electoral political speech versus a post-electoral press-conference.

Following Lindblom's H&H theory [5], we should say that this particular contextual opposition optimized a dichotomy between one context – here the pre-electoral political speech – where output constraints dominate and hyperforms are expected and a another context – here the post-electoral press-conference – where system contraints dominate and hypoforms are selected. The results show that pitch range and pitch register obviously contrast along the hyper-hypospeech dimension as defined by Lindblom.

CONTEXTS, METHODOLOGY AND RANGE & REGISTER

There is without doubt a current interest in investigating spontaneous speech. However, the term 'spontaneous speech' covers a considerable range of speech corpora. In this respect, it might be important to establish a difference between a corpus of spontaneous speech elicited *intra muros*, in the context of a phonetic laboratory for experimental purposes and a corpus of spontaneous speech produced *extra muros*, in the context of social interaction and without any experimental purpose. Why not call the former 'spontaneous lab speech' and the latter 'spontaneous speech'? (for a general discussion concerning phonetics and real speech, see [6], for the contrast between spontaneous 'lab' speech and read 'lab' speech see [7] and for an interesting attempt to simulate rhetoric in the lab see [8]).

We would like to suggest that this terminological adjustment seems revelant also insofar as it has impact on data collection, experimental methodology, and modelling paradigms.

Choosing contexts

In controlled laboratory experiments, the researcher's experimental setting can create contextual frames that are not always consistent with the informant speaker's everyday practices. This is hardly the case with *extra muros* conditions, where the context is de facto given by the social interaction. The choice of relevant contexts is therefore crucial when collecting spontaneous within-speaker data under *extra muros* conditions.

Here, the opposition pre-electoral versus post-electoral speech has been chosen because this is when persuasion (when a politician aims to gain votes) gives way to a non-persuasive pathos (when he comments his political victory or defeat). A basic assumption underlying the choice of the broadcasted pre-electoral speech is that every achievement in the referential field directed to the listening public must in reality be translated into conative achievements in the politian effort to persuade listener-voters. We can therefore assume that a pre-electoral context shaped the speaker's speech toward 'hyperform' speech, as opposed to a post-electoral context which shaped the speaker's speech toward 'hypoform' speech.

Methodology

Because we were working with spontaneous speech, the question of the research setting proved to be essential. It has resulted in a methodology where restricted samples of speech material – short and well time-defined discourse events – from conversations, interviews, political debates, political speeches and radio programs have been studied from different angles.

We have conducted four different kinds of analyses (see [9]): (1) analysis of the discourse structure of the speech corpus without specific reference to prosodic information (in order to avoid circularity), (2) auditory analysis (which is implemented as a selective prosodic transcription in Waves+), (3) acoustic-phonetic analysis, and (4) analysis-by-synthesis.

By first focussing our attention on general discourse characteristics on a speech fragment produced in a particular context and by then following a rather classic phonetic analysis, we are meeting basic methodological requirements that allow us to propose a formalized prosodic description of successive individual utterances in their specific *extra muros* context.

Modelling range & register

The majority of studies dedicated to the analysis and modelling of range and register are carried out within the *intra muros* condition where the speaker is asked to read isolate sentences simulating different emotional states or attitudes. In [10], Bruce showed that differences in attitude (detached-involved) involve pitch range variation, achieved by Fo expansion upward, the lower Fo limit being fixed. Even if Bruce noticed that not only the maxima were raised, but the minima also, he was not interpreting these raised Fo minima as changes in local register. On the other hand, Gårding's tonal grid parameters [11] clearly proposed two different types of Fo expansion between parallel lines. One denoted as 'R' which expressed a global range and another, 'r', which is the vertical distance between the grid lines which might be interpreted as change in register. Similar parameters were used by Ladd for his CSTR model [12]. The main difference is that for Ladd register makes reference to target level and not to shape as is the case for Gårding. More prosaically, Swerts & Collier [13] defined register as the mean Fo of a speech fragment (expressed in Hz) and range as the standard deviation from the mean Fo of the speech fragment (expressed in semitones).

Overall pitch variation in spontaneous French speech fragments collected under *extra muros* conditions was observed by Mertens [14]. In order to categorize these changes, Mertens proposed three registers for French: a middle register, a low register and a high register. The middle register is placed in the central part of the speaker's tonal range – it constitues the speaker's usual register. Changes from this central tonal register toward a lower or higher tonal register imply new values for the Fo interpretation of the High and Low turning-points. However, our analysis of overall pitch within the specific setting of a pre-electoral speech [4] provided evidence that we needed a two-fold categorization of overall pitch variation in French – one in terms of range and the other in terms of register. I also proposed an adjustment of the KIPROS transcription system with regard to overall pitch.

PROCEDURE

The recorded material were digitized and analyzed using the ESPS/Waves+ environment which enables transcription and labelling in multiple tiers.

An important step was the auditory analysis which provided an orthographic and a prosodic transcription of what had been recorded. More specifically, prosodic features marked for the purpose of this experiment were phrasing and overall pitch i.e. range and register.

For each transcribed prosodic phrase, a statistical program (see [15]) performs calculations on the Fo file. Fo values were collected for three pitch parameters: a local absolute Fo minimum (Fo_{min} with its temporal location), a local absolute Fo maximum (Fo_{max} with its temporal location), and a global parameter that is an average Fo (Fo_{mean}) over the whole phrase. These values were then directed to a new file and could be viewed for control with xlabel (ESPS). Detected Fo points with erratic values could be assigned manually to a new temporal location.

The values obtained for successive prosodic phrases in each setting were plotted as presented in Figures 1 and 2. All the values were also pooled and presented in Table 1. In order to capture differences in Fo variation between the different contexts, frequency modulation factors (SD/mean in %) as proposed in [17] were also calculated.

OBSERVATIONS

Figures 1 and 2 show Fo_{max}, Fo_{mean} and Fo_{min} values (with average) for successive prosodic phrases produced in the two different contexts. For the three parameters, the Fo variation is larger in the first context than in the second context. In Fig.1 the prosodic phrase 4 is representative of a reduced range but a high register. On the contrary, the prosodic phrase 10 is a good example of an expanded range with a relatively low register.

Table 1 showed that all values for Fo_{mean}, Fo_{min}, and Fo_{max} are systematically higher in the pre-electoral speech. However, the Fo expansion (calculated as Fo_{max}/Fo_{min}) used by the speaker in the two contexts is very similar in proportion across the two contexts (2.2 for the first and 2.1 for the second).

Frequency modulation factors pointed out a particular increase in variation for Fo_{max} in the first context. But it is worth noting the absolute high level of Fo_{min} which indicates also the use of high register in this context.

The Fo_{min} (91 Hz) in the non-emphatic post-electoral speech seemed to serve as a default base value (it is very near the base-value (Fb=93.4) for a male speaker of European languages (see [16])

Table 1. *Average Fo (means and standard deviation; values are in Hz) for Fo_{mean}, Fo_{min}, Fo_{max}, and frequency modulation factors in two different contexts (A: pre-electoral speech, B: post-electoral press-conference).*

	A	B
Fo_{mean}	229.1	139.4
SD	32.8	18.1
SD/mean %	14.3	12.9
Fo_{min}	141.9	91.0
SD	39.5	19.3
SD/mean %	27.8	21.2
Fo_{max}	317 .7	199.7
SD	69.1	19.6
SD/mean %	21.7	9.8

CONCLUDING REMARKS

To conclude, I would like to define two kinds of range: a voice's range which is the Fo distance between the absolute Fo maximum and the speaker's baseline (specified as the usual Fo floor), both reached across contexts, and a context specific range which is the Fo distance between the absolute Fo maximum and the speaker's baseline (specified as the Fo minimum in a specific context). A register would be defined in terms of Fo level given by absolute Fo minimun in actual prosodic phrases.

Obviously, this politician used a bi-modal overall Fo distribution, high-pitched in the pre-electoral speech and relatively low-pitched in the post-electoral conference. The pre-electoral speech, in contrast with the post-electoral speech, seemed to shape the speaker's intonation toward a more hyperform behaviour with larger and more variable Fo excursions and several changes in register.

ACKNOWLEDGEMENTS

I am especially grateful to Marcus Filipsson for his computational assistance.

REFERENCES

[1] Touati, P. (1991), "Temporal profiles and tonal configurations in French political speech", *Working Papers*, vol. 38, pp. 205–219.
[2] Duez, D. (1991), *La pause dans la parole de l'homme politique*, Paris: Editions du CNRS.

[3] Bhatt, P. and Léon, P. (1991), "Melodic patterns in three types of radio discourse'", *Proceedings of the ETRW 'Phonetics and Phonology of Speaking Styles'*, Barcelona, pp. 11:1-11:5.
[4] Touati, P. (1993), "Prosodic aspects of rhetorical prosody", *Working Papers*, vol. 41, pp. 168-171.
[5] Lindblom, B. (1990), "Explaining phonetic variation: A sketch of the H&H theory", in *Speech Production and Speech Modelling*, Eds. Hardcastle, W. and Marchal, A., Kluwer Academic Publishers, pp. 403-439.
[6] Rischel, J. (1992), "Formal linguistics and real speech", *Speech Communication*, vol.11, pp. 379-392.
[7] Bruce, G. (forthcoming), "Modelling Swedish intonation for read and spontaneous speech", *Proceedings of the XIII ICPhS*, Stockholm.
[8] Gussenhoven, C. (1983), "Stress shift i Dutch as a rhetorical device", *Linguistics*, vol. 21, pp. 603-619.
[9] Bruce, G. and Touati, P. (1992), "On the analysis of prosody in spontaneous speech with exemplification from Swedish and French", *Speech Communication*, vol. 11, pp. 453–458.
[10] Bruce, G. (1982) "Developing the Swedish intonation model", *Working Papers*, vol. 22, pp. 51-116.
[11] Gårding, E. (1991), "Intonation parameters in production and perception", *Proceedings of the XII ICPhS*, Aix-en-Provence, pp. 300-304
[12] Ladd, R. (1990), "Metrical representation of pitch register", *Papers in Laboratory Phonology I*, Eds. Kingston, J. and Beckman, M., Cambridge University Press, pp. 35-57.
[13] Swerts, M. and Collier, R. (1992), "On the controlled elicitation of spontaneous speech", *Speech Communication*, vol. 11, pp. 463-468.
[14] Mertens, P. (1987), *L'intonation du français. De la description linguistique à la reconnaissance automatique*, Katholieke Universiteit Leuven.
[15] Filipsson, M. (1995), LABF0_STATS, Dept. of Linguistics, Lund University.
[16] Traunmüller, H. and Eriksson, A. (1994), "The frequency range of the voice fundamental in the speech of male and female adults.", Manuscript. Inst. of Linguistics, University of Stockholm.

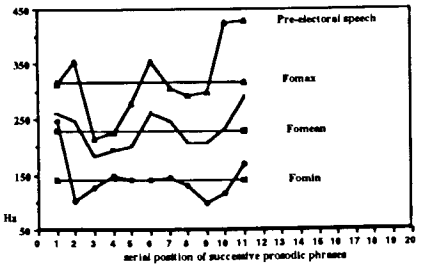


Figure 1. Fo_{mean}, Fo_{min}, Fo_{max} (with average) in pre-electoral speech (values are in Hz).

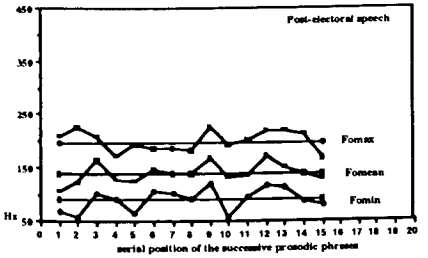


Figure 2. Fo_{mean}, Fo_{min}, Fo_{max} (with average) in post-electoral press-conference; values are in Hz

USING SEGMENTAL DURATION PREDICTION FOR RESCORING THE N-BEST SOLUTION IN SPEECH RECOGNITION

K. Bartkova, D. Jouvet & T. Moudenc
CNET, France Télécom, Lannion, France

ABSTRACT

The aim of this study is to set up a rule-based model of segmental duration in French for an automatic speech recognition system. This model was introduced at the post-processing stage of speech recognition in order to rescore the N-best solution hypotheses. In preliminary experiments conducted on isolated and connected word databases, the reduction in the recognition error rate ranged from 11 % (for numbers) to 19 % (for digits) when duration information was used in post-processing.

INTRODUCTION

Prosodic features contain valuable information about speech structuring. A great number of studies have focused their attention on the measure and description of principal physical prosodic parameters such as F0, sound duration and sound intensity. Researchers in automatic speech processing have long been aware of the importance of integrating prosody into different automatic systems. In speech synthesis the role of prosody is clearer: it *must* be modelled for generating natural sounding speech. In automatic speech recognition, prosodic parameters have primarily been used in order to segment signals into prosodic units [1]. The main prosodic cues used for signal segmentation are: final syllable lengthening of a prosodic constituent, and F0 movement amplitude.

A great number of studies deal with sound duration modelling in recognition systems based on Hidden Markov Modelling (HMM). Some of these studies use minimal sound duration [2], or sound duration normalized by utterance length [3], or variable sound duration according to speech rate [4].

THE RULE-BASED MODEL

The aim of this study is to introduce phonetic knowledge into sound duration modelling in order to set up a phonetically-based sound duration model. This phonetic duration model is then used in the post-processing of the N-best solutions hypotheses given by an HMM based system using only spectral representation information. The role of the duration prediction is to distinguish between good and bad solutions proposed by the recognizer. Thus, either the duration score confirms the scoring of the HMM, or, conversely, penalizes the score (for example, when the duration of the segments constituting the solution does not match the duration predicted by the model).

The rule-based duration model for French sounds predicts segmental duration according to relevant phonetic and phonologic events. Phonemes are grouped into macro-classes. There are 4 vocalic macro classes (oral vowels, nasal vowels, neutral schwa-like vowels and semi-vowels), and 7 consonantal macro classes (voiceless plosives, voiced plosives, voiceless fricatives, voiced fricatives, nasals, r and l). For each macro class, mean phoneme durations and standard deviations were calculated according to: the left and right context (also expressed in macro classes), the word length, and the position of the syllable in the word (final syllable versus non-final syllable).

CONTEXT GROUPING

Several studies in micro-prosody illustrate that right consonant contexts have greater influence on vowel duration than left consonant contexts[5]. In French, the accent falls on the last syllable of a prosodic unit. As such, the vowel duration is clearly dependent upon the right context only in final "stressed" syllables. In order to obtain appropriate phonetic modelling, considering the right context of the last syllable of a lexical word is sufficient. Unfortunately, HMM segmentation is not always accurate. A type of "spectral inertia" persists in the segmentation process. This is due to the fact that parameters used for modelling contain mainly spectral information (8 Mel Frequency Cepstral coefficients and their first and second order temporal derivatives), and only three values related to energy (energy value and its first and second order derivatives). One can assume that mistakes made by HMM in segmentation are consistent for this very reason. In order to surmount this segmentation defect, both left and right contexts are taken into account in the sound duration modelling. In terms of syllable vowel duration, when a consonantal cluster closes the syllable, in addition to the immediate right consonantal context, the last consonant is also taken into account. This is because the phonetic characteristics of the last consonant (together with syllable structure) influence vowel duration.

PARAMETER SMOOTHING

Smoothing was implemented when the number of occurrences was not high enough to enable the reliable estimation of a sound duration parameter. Among the different phonetic parameters an *a priori* hierarchy was established, which specifies the order in which the smoothing is conducted. During rule smoothing, all the other parameters remain unchanged. For example, for a vowel, the first phonetic parameter to be considered is the left context, while the other conditions (right context, syllable position, word length) stay unchanged. Figure 1 shows the hierarchical clustering used in the smoothing of the left context. One moves up the tree until a sufficient number of occurrences is found; the corresponding parameters are then judged reliable.

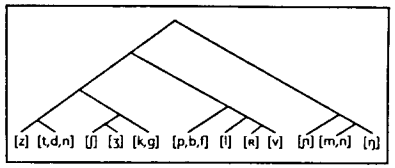


Figure 1: Left Consonant Context Clustering for Vowels

Other phonetic parameters are considered (left context, etc.), and ultimately the inherent sound duration is used. However, the inherent duration is shortened for non-final syllable positions.

Initially, sound duration was modelled for 3 different corpora in French: digits, numbers from 00 to 99, and 36 words and expressions (Trégor corpus). All three corpora were recorded through the telephone network using about 800 speakers. Half of the data was used to train the predictive models, and the other half was used for testing. The allophone units were modelled by HMM [6].

Although the corpus contained connected words, the relative shortness of the sentences (the longest one containing 7 syllables), and the poverty of their syntactic structure (belonging to the same constituent), prevented the prediction of the sound duration in the final syllables in terms of the different depths of the syntactic structure of the sentence. Since internal pauses (if any) inserted between two adjacent words were always relatively short, and their duration quite consistent, it proved useful to model these durations too.

MODEL PARAMETERS

Two different models were set up using the three corpora described herein. Two training procedures were evaluated: one was corpus-dependent, and one was pluri-corpus. Corpus-dependent means that the sound duration parameters were trained and used on the same corpus

(same vocabulary but different utterances). In the pluri-corpus training, the sound duration parameters were trained on the 3 corpora combined, and then used on each of them individually. The second model contained a more detailed context definition (16 macro classes instead of 5, as in the first one) in order to compensate for the defects of the spectral inertia. A third, corpus-independent model, was trained using hand-segmented data which differed from the corpora in this study. The best performance was obtained using the pluri-corpus trained model which contained a refined context definition, subsequently only these results will be analyzed herein.

One of the principal characteristics of duration is elasticity. Some speakers articulate faster than others, and the same speaker can change speech rates at any time, even during the same sentence. In order to deal with speaking rate phenomena, two speaking rate coefficients were calculated for each sentence: one for consonants, and one for vowels. These coefficients minimize the global error between predicted and measured duration. The resulting segmental duration errors (between measured and predicted duration) were modelled separately for correct and incorrect alignments. These two models were incorporated into the post-processing in order to rescore the N-best solutions. Figure 2 shows prediction duration errors for French digits.

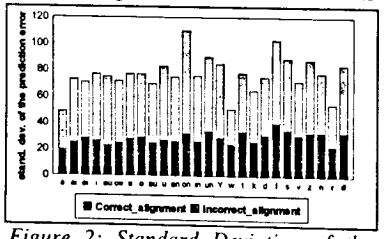


Figure 2: Standard Deviation of the Sound Duration Prediction Errors

The prediction error is minimized, and its value approaches 0, using speech rate coefficients in a monosyllabic word containing only one consonant and one vowel (as in the digit "deux" in French). If only error prediction is modelled, incorrect alignments associated with monosyllabic models cannot be penalized. Thus, speech rate coefficients must also be modelled. Figure 3 illustrates the consonant speaking rate histogram for the digit corpus .

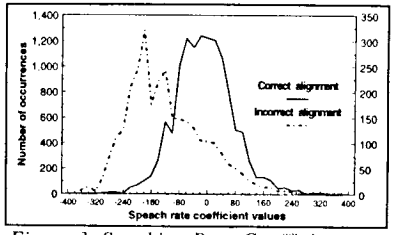


Figure 3: Speaking Rate Coefficients

MODEL EVALUATION

The efficiency of the sound duration predictive model was evaluated in the post-processing procedure. The benefits of different components (e.g., prediction error and speech rate coefficients), were initially tested separately, then together, and finally in combination with HMM scores.

Detailed results show tha this processing proved to be beneficial, as did the possibility of recovering HMM errors with duration prediction. Table I provides correct recognition percentages using duration and speech rate information, separately and combined. The percentage of the HMM errors recovered by duration post-processing was about 50% for each corpus, duration and speech rate scores combined.

Table I : Percentage of Correct Identification using only Duration Information

	Digits	Numbers	Trégor
Duration prediction error	**45 %**	**62 %**	**64 %**
Speaking rate	**68 %**	**38 %**	**51 %**
Duration prediction error + Speaking rate	**76 %**	**67 %**	**80 %**

Table II: Test Set Recognition Error Rate Reduction for Several Post-processings

	Digits	Numbers	Trégor
HMM+Duration error+Speaking rate	5 %	8 %	8 %
HMM+Duration error+Stationarity	17 %	17 %	6 %
HMM+Speaking rate+Stationarity	17 %	7 %	2 %
HMM+Duration error+Speaking rate+Stationarity	**17 %**	**16 %**	**8 %**

As illustrated in Table I, speech rate modelling works well for short, mainly monosyllabic vocabularies, such as digits. Duration modelling works better for longer word vocabularies (such as Numbers or the Trégor). Regardless, the combination of both scores provides relatively good results, considering that only duration information was used.

Table II illustrates the recognition error rate reduction, following the introduction of the duration prediction error score and/or the speech rate score in post-processing, in comparison with HMM alone. Preliminary tests were conducted, recombining duration information and a supplementary parameter obtained from an *a priori* segmentation of the speech signal (this parameter expresses the number of stationary zones that occur in each segment of the signal) [6].

CONCLUSION AND DISCUSSION

This study focuses on the investigation of new parameters used in speech recognition system post-processing. It is reasonable to assume that the introduction of different types of parameters can add valuable information to the rescoring of the HMM spectral score, where scoring is achieved using speech spectral representation following an initial pass through the system. Sound duration rule-based prediction is one type of supplementary parameter. Although information supplied by sound duration is poorer than those of spectral word representations (two different words or hypotheses can have exactly the same duration), initial attempts at evaluating the efficiency of these parameters have proved quite hopeful.

Further attempts to introduce other prosodic parameters such as F0, in order to rescore the N-best solutions in post-processing, are also being made.

REFERENCES

[1] Vaissière, J. (1989), "On automatic extraction of prosodic information for automatic speech recognition system", *EUROSPEECH'89,* Paris, pp. 26-30.

[2] Gupta, V. Lenning, M. Mermelstein, P. Kenny, P. Seitz, P.F. & O'Shaughnessy D. (1992), "Use of minimum duration and energy contour for phonemes to improve large vocabulary isolated-word recognition", *Computer Speech and Language,* vol 6, pp. 345-359.

[3] Gong, Y & Treurniet C. W. (1993), "Duration of phones as function of utterance length and its use in automatic speech recognition", *EUROSPEECH'93,* Berlin, pp. 315-318.

[4] Suaudeau, N. & André-Obrecht R. (1993), "Sound duration modelling and time-variable speaking rate in a speech recognition system", *EUROSPEECH'93,* Berlin, 307-310.

[5] Di Christo, A. (1978), *"De la prosodie à l'intonosyntaxe",* Thèse de Doctorat d'Etat, Université de Provence, Aix-Marseille I.

[5] Bartkova, K. & Jouvet D. (1991), "Modelization of allophones in a speech recognition system", *ICPhS'91,* Aix-en-Provence, pp. 474-477;

[6] Moudenc, T. Jouvet; D & Monné, J. (1995), "On using a priori segmentation of the speech signal in N-best solutions post-processing", *ICASSP'95,* Michigan, USA.

A FLEXIBLE VOCABULARY RECOGNITION SYSTEM FOR ITALIAN

P. Bonaventura Δ, L. Fissore, H. Leprieur ◊ and G. Micca
CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - 10148 Torino, Italy
Δ CSELT Consultant ◊ PhD student

ABSTRACT

The present paper describes a flexible vocabulary speech recognition system developed at Cselt. Main modules are described and some recognition results are provided. Details on the phonetic component are given. Preliminary tests have been performed on a speaker-independent, continuous speech recognition task, using the most recent release of the recognizer.

INTRODUCTION

CSELT research in the field of automatic speech recognition is presently being carried out along two main directions: a) development of real-word applications based on speech technologies already mature, as Voice Dialling by name [1] and advanced research activities for speech understanding through natural language in man-machine interaction systems with spontaneous dialogue [2].

One key factor in the advancement of speech recognition technology was the shift from Whole Word modeling to sub-word modeling technology. This latter, adds the important feature of Vocabulary Independency to speech recognition systems and paves the way to the wide range of flexible vocabulary applications [3].

Since September 1994 CSELT is internally experimenting a voice dialling-by-name service based on FLEXUS®, a flexible vocabulary recognizer trained in speaker-independent mode for the telephone network. Currently the application supports nearly 1,000 surnames and runs in nearly real time on a Personal Computer equipped with a multichannel DSP board. The clear separation between Language-Dependent and -Independent components, as well as the close modularization of the architecture makes the system easily extensible to other languages.

SYSTEM OVERVIEW

The block diagram of the speech recognizer is depicted in Fig1.

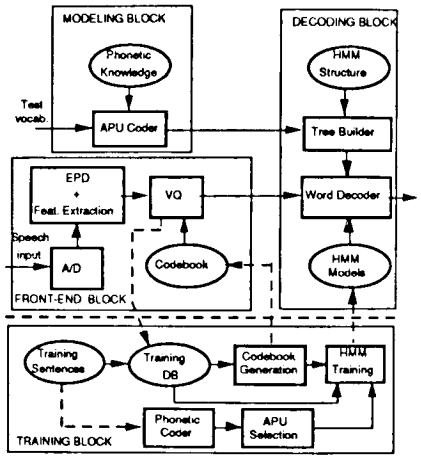


Fig.1 *Block diagram of the speech recognition system*

The training phase is carried out off-line on a SUN Workstation. The **Feature Extraction** module computes spectral features from 10 ms segments (frames) of digitized speech, then a **Vector Quantization** module computes codeword indexes for each input frame with respect to a given codebook. The **Phonetic Coder** transcribes the lexicon extracted from the whole corpus of training in terms of phonemes, then an appropriate set of **Acoustic-Phonetic Units** (APU's) is generated which suitably satisfies two fundamental but competing requirements: *precision* of the model, that is higher with a larger number of units, and *trainability*, that is greater with a reduced number of units. APU's are then associated to Hidden Markov Automata with a given topological structure. The **HMM Training** module learns a stochastic structure for each state of the model (transition and emission probabilities). The on-line decoder consists of a **Lexical Tree Builder** which represents the target lexicon in terms of a tree of sub-word units, of a **real time** version of the Feature Extractor and VQ modules, and of a real time implementation of the **Word Decoder** which computes likelihood scores for word or sentence candidates, sorts them and provides the best one or the N-best ones in output.

LMS

The LMS (Lexical Modeling System) component provides the lexical and acoustic-phonetic knowledge to the recognition system and is divided in modules communicating through simple data structures. Both declarative and procedural knowledge is included. The prominent declarative knowledge is represented by the phonetic coding rules, by the phonetic and APU tables and by coarticulations rules. Procedural knowledge consists of algorithms for sentence generation, lexicon extraction, phonetic rule instantiation, APU generation, graph representation and statistics computations. Phonetic rules and tables, prototypes and semantic classes can be easily substituted for those of a new language, giving LMS a multilingual processing capability. Presently, an English version is being experimented within the RAILTEL EEC-funded project.

Syntactic prototypes

Sentences in a given application domain can be automatically generated from syntactic prototypes represented by graphs. Arcs in a graph are associated to *terminal symbols* or to *semantic classes*. Terminal symbols are single words or sequences of words. Phonetic classes are sets of words associated to a semantic concept: f.i., <TOWNS>, <NUMBERS>, <COMPANIES>. Syntactic forms can also be grouped in classes: *"vorrei", "potrei", "desidererei"* are members of the class <QUERY-START>. Sentences are then generated by visiting the graph along randomly chosen different paths.

Phonetic rules description

The Phonetic Coder includes grapheme-to-phoneme conversion rules for transcription of standard italian and of major regional pronunciations. The regional rules describe the most common linguistic phenomena (assimilation, insertions and deletions) that modify standard Italian when pronounced according to a regional variety [6]. The regional rules have been subdivided into three categories, according to their distribution on the national territory: category 1 includes pronunciations present at least in 5 regions, category 2 describes variants present in three to two regions, and category 3 includes pronunciations typical of one region.

Thirteen regional variants [6] have been selected as representative of the twenty official Italian dialects: Piedmont (PI), Liguria (LI), Lombardy (LO), Veneto (VE), Giulia (GI), Emilia (EM), Sardinia (SA), Tuscany (TU), Umbria-Marche (U-M), Lazio (LA), Apulia (PU), Campania (CA) and Sicily (SI). The regional rules have been written and implemented in the SCYLA language [5], whose syntax is very similar to the generative one; the alphabet used for the formulation of the regional rules includes the following elements:

{ }= 'aut' (exclusive 'or')
C= any consonant
V= any vowel
##= word boundary
+= morpheme boundary

An example of rewriting rule is the following:

$$s \rightarrow \int / \underset{[-\text{voiced}, +\text{stop}]}{\_\_} C$$

(SICILIA, Ex. 'aspetto')

The features adopted (e.g. 'voiced', 'stop') define the phones on the basis of phonetic categories (place and mode of articulation), and do not have phonological value. The SCYLA grapheme-to-phoneme transcription component converts an input text string into an output phonetic transcription. Rules for standard Italian [10] have already been written in SCYLA and implemented in a text-to-speech synthesis.

The transcription rules actually convert graphemes into phones and not phonemes: however, this linguistically

inexact terminology has been used throughout the paper, because it is conventionally accepted in the field of speech synthesis and recognition.

Linguistic phenomena described by the rules

The rules have been subdivided into three categories: Category 1 rules describe phenomena which occur in 7 regions (usually the northern ones vs. center-southern ones). Linguistic phenomena common to these areas are: a) voicing of intervocalic /s/ (e.g. 'casa', 'house': central-south. ['kasa] vs. north. ['kaza]). b) pronunciation of [s] as a dental affricate [ts], after a nasal and a liquid (e.g. 'arso', 'ansa' and 'salsa': stand. ['arso], ['ansa], ['salsa], regional ['artso], ['antsa] and ['saltsa]); c) pronunciation of velar allophone [ŋ] before consonant in the north of Italy (e. g. 'cantare': standard [kan'tare], regional [kaŋ'tare]); d) insertion of glide [j] after palatal affricate, liquid and nasal consonants [tʃ], [ʃ], [ʎ], [ɲ] if corresponding graphemic groups 'ci , 'sci', 'gli', 'gni' precede a vowel (e.g. 'cielo': standard ['tʃɛlo], regional ['tʃjɛlo]; 'scienza': standard ['ʃɛntsa], regional ['ʃjɛntsa]; 'gliene': standard ['ʎene], regional ['ʎjene]; 'sogniamo': standard [soɲ'ɲamo], regional [so'ɲjamo]).

Category 2 rules are present in 3 to 2 regions and describe the following phenomena: a) voicing of the affricate [ts] before vowel and after [n], in TO, PU, CA (e.g. 'fidanzato': stand. [fidan'tsato], region. [fidan'dzato]; b) doubling of intervocalic [b] and [dʒ] in SI, CA, LA (e.g. 'cabina': stand. [ka'bina], region. [kab'bina]; 'agiato': stand. [a'dʒato], region. [ad'dʒato]; c) progressive assmilation of consonant clusters [tm], [kn], [nr] in TO, U-M, LA (e. g. 'atmosfera': stand. [atmos'fɛra], region. [ammos'fɛra]; 'tecnico': stand. ['tɛkniko], region. ['tɛnniko]; 'Enrico': stand. [en'ri: ko], region. [er'riko]; d) simplification of the intervocalic palatal lateral [ʎ] as the glide [j] in U-M, TO (e.g. 'moglie': stand. ['moʎ:e], region. ['mojje].

Linguistic phenomena present in one region (category 3) are the following: a) palatalization of [s] in EM in every context (e.g. 'sposare': stand. [spo'sare], region. [ʃpo'ʃare]; b) realization of [l] as [r] before consonant in Lazio (e.g. 'alzare': stand. [al'tsare], region. [ar'tsare]).

APU generation

The set of APU's is selected as a viable trade-off between precision of the model and statistical robustness. Acoustic variability can be better captured by differentiating phones with respect to their context: triphones specify a left and a right context, biphones a left or a right context. *(t) r (e)*, f.i., is phone [r] in the word 'treno'.

Monosyllabic functional words (articles, prepositions), which are more prone to heavy coarticulation in continuous speech, are given specific, word-dependent units. The large number of possible triphones in Italian (a few thousands) is downsized by a) imposing a minimum occurrence threshold and b) backing off to biphones.

Graph concatenation

APU's are chained up to form whole words, and words are chained up to form sentences. Therefore, at the end of the concatenation process a single sentence is represented by a graph, which includes pronunciations variants, optional silences and schwa's, and allophonic alternatives. A two step optimization was implemented: first, linear sequences of APU's are obtained for each phonetic variant of every word, then a first optimization is carried out to obtain a word graph; successively, the word graph is chained to the graph corresponding to the sub-sentence already processed, and a second optimization is performed to minimize the number of nodes of the overall graph. The complexity of a sentence graph is ranges from the simplest level (only standard transcription) to the most complex level (all regional variations included).

APU coder

This module codes a given test vocabulary in terms of the APU's alphabet. Linear representations of word are obtained, which will be given in input to the Word Tree Builder to provide a compact tree structure for the on-line Decoder.

HMM topology

Each unit is given a topological structure, based on Bakis' linear model with loop and forward arcs. In the standard three state model, lateral states account for coarticulation effects with adjacent sounds, while the central state captures the stationary component of the phoneme.

Coverage

Statistical tools are included in LMS, to provide general information about the distribution of words, phones, APU's in lexical corpora. The coverage parameters are computed as percentage of occurrence of a given type of APU within a corpus. The coverage parameters are important in evaluating the level of precision of a given model with respect to a training corpus, and the degree of "generalization" of that model with respect to a new application lexicon.

TESTS

Experimental Set up

The speech signal is collected through a linear electret-type telephone connected to a local PBX, filtered in the telephone bandwidth of 300-3400 Kz, and sampled at 8 Khz. 12 Mel-based cepstral and Δcepstral coefficients are computed each 10 ms time frame, plus Energy and ΔEnergy. A variable threshold, energy-based end-point detector is used. Two 256-codeword codebooks are generated for Cepstrum and ΔCepstrum, plus a 32-codeword one for Energy and DEnergy. Discrete density HMM's were trained through a few Forward-Backward training iterations. A beam-search Viterbi decoder generates the best-scored sequence of words along which we compute the Word Accuracy by alignement with the pronounced sequence [4].

Performance results

Preliminary results were obtained by training the recognizer on a 12,000 utterances data base from 146 speakers, and by testing on 600 utterances from 10 diferent speakers, in the train information query domain, with a vocabulary og 718 words. Utterances were read on a screen in a quiet room, with a linear (electret) microphone, on the local PBX. A Word Accuracy figure of 76.6% has been obtained with Discrete density HMM's, which increased to 79% by using Continuous density HMM's, and to about 90% with statistical language models (word bigrams). Present experimentation aims to improve recognition performance by adopting new types of APU's, including transitional and stationary components, to adapt the recognizer to new languages and to increase the robustness of the models to channel and line variations. The capability of the acoustic-phonetic module to match specific regional pronunciations will be assessed in future experiments.

CONCLUSIONS

A description of the Flexible Vocabulary Recognition Technology developed at Cselt has been given. The most important features of the new release of the LMS component for lexical and phonetical processing have been described. Some recognition tests have been carried out on a speaker-independent, continuous speech recognition task over the telephone, and the corresponding performance figures have been presented. Finally, future developments are mentioned.

REFERENCES

[1] Ciaramella A., Clementino D., Fissore L., Pacifici R., Sperti S.(1993) *"Voice Dialling by name in a PBX environment"*, ESCA Workshop, Bavaria, Germany, pp. 179-182.

[2] Baggia P., Fissore L., Micca G., Rullent C., Laface P.(1994) *"A Speech Understanding System for Information Retrieval"* Int. J. Patter. Rec. and A.I., Vol. 8(1), pp. 71-97.

[3] Lennig M. et al. *"Flexible Vocabulary recognition of speech"* ICSLP'92, Banff (Canada), pp. 93-96.

[4] Fissore L., Laface P., Micca G. (1991)*"Comparison of Discrete and Continuous HMMs in a CSR Task over the Telephone"*, ICASSP '91, Toronto, pp. 253-256.

[5] Nebbia L., Lazzaretto S.(1987) *"SCYLA: Speech Compiler for Your LAnguage"*, Europ. Conf. Speech Techn., Edinburgh.

[6] Canepari L. (1979), *"Introduzione alla fonetica"*; Torino, Einaudi.

[7] Salza P.G.(1990), *"Phonetic Transcription rules for Text-to-Speech Synthesis of Italian"*; Phonetica, (47), pp.66-83.

AUTOMATIC SPEECH RECOGNITION USING PRODUCTION MODELS

Laurence CANDILLE, Henri MELONI

Laboratoire d'informatique
Faculté des Sciences, 33 rue L. Pasteur, 84000 avignon
tel.: 90 14 44 21
e-mail: candille@univ-avignon.fr

ABSTRACT

We present how the Distinctive Region Model (DRM) may be used for the recognition of two vowel sequences. The process studied here takes into account the characteristics of the speaker, the phonetic context and the variation of the formants during the V1-V2 transition. On average, 80% of the V1-V2 sequences are correctly identified. The article presents the results obtained for the ten French vocalic vowels.

1- INTRODUCTION

In order to verify whether speech production models are appropriate for automatic speech recognition, we primarily use the DRM model because it is simple and easy to control.

This model offers the advantage of proposing a dynamic modelisation of articulators motion to go from one vocal tract configuration to another. Moreover, it is speaker independent. The model parameters are the areas of the regions and the total length of the tube.

In this preliminary stage, we consider the identification of V1-V2 sequences uttered by several speakers. The model derived formant transition are compared with the acoustically measured ones.

2- DRM MODEL INVERSION

The aim of the present work is not to validate the model nor to modify it to solve particular difficulties. We use it as it has been designed by its conceptors [2]. The DRM model inversion process and our recognition strategy have been described in [1].

2.1 - speaker adaptation

In order for the model acoustic space to better match the speaker's, we must either modify some characteristics of the model or normalise the speaker's parameters.

The DRM model characteristics allow a speaker adaptation by varying the total tube length.

The model may be adapted in two ways: either by fixing the total vocal tract (VT) length for each speaker and is identical for every vowel or by fixing the length for each vowel of each speaker.

2.2 - codebook generation

In order to optimize the static search for configurations which constitutes the first part of our recognition strategy, we generate a codebook i.e a table of acoustic vectors and corresponding articulatory vectors which provides starting and final configurations for each transition.

The configurations in the codebook are produced by varying around a reference VT model. There is a reference vocal tract model for each vowel. The variation allowed around each reference configuration is fixed. All other configurations are produced by moving from one extreme configuration to another along a straight line in the parameter space. We also vary the interpolation type and the configuration total length.

Thus we obtain a reference table (TR -0) containing about 15 000 configurations which describe the whole vowel set.

Some VT models are associated with acoustic parameters which do not match those of the currently studied vowel, therefore they must be filtered out. By filtering the table TR-0, we create six different tables distinct from each other by the number of configurations for each vowel, the choice of these configurations and their total length. These tables will be used directly for the recognition of the V1-V2 sequences. The first working-table TT-1-1 contains the configurations of the reference table TR-0. The total length of the configurations is fixed for each speaker and is identical for each vowel. TT-1-2 contains 20 configurations per vowel (from TR-0); these configurations are speaker dependent and represent every vowel in context. Finally in table TT-1 -3 the configurations are speaker dependent and each vowel is represented by one configuration only.

The last three tables TT-2-1, TT-2-2, TT-2-3 are built respectively like the first three ones but the configurations length is fixed for every vowel of every speaker.

2.3 - recognition strategy

Firstly we measure the first three formants at the onset and the offset of the input signal. Refering to the codebook, we determine a VT configuration for each hypothesis concerning V1 and V2. All possible formant transitions are then calculated using the speaker adapted model with its two commands and different interpolation types. Secondly we identify the V1V2 sequences whose entire formant transition matches best the formant transition measured on the input signal.

3- RESULTS

3.1 - symetric model

Three male French speakers uttered a hundred of V1-V2 sequences consisting of the ten French oral vowels.

The standard DRM model configurations, even with varying lengths, yield no more than 50% recognition rate on average for the three speakers (these results must be compared with tables TT-1-3 and TT-2-3).

The recognition rate in first position of the V1-V2 transitions, for each speaker, and for each working-table is stated in Tab1.

Tab 1: The recognition rate in first position of the V1-V2 transitions, for each speaker (across), and for each working-table (down). The first three tables TT-1-1, TT-1-2, TT-1-3 contain respectively about one hundred configurations, 20 configurations and 10. The total length of the configurations is fixed for each speaker, and is identical for each vowel. The last three tables TT-2-1, TT-2-2, TT-2-3 are built respectively like the first three ones but the configuration length is fixed for every vowel of every speaker.

	TT 1-1	TT 1-2	TT 1-3	TT 2-1	TT 2-2	TT 2-3
TS	52%	67%	69%	65%	80%	80%
FB	42%	56%	55%	54%	77%	80%
PG	50%	65%	68%	50%	79%	81%
TOT	48%	63%	64%	56%	79%	80%

These first results show that the VT model length adaptation for each speaker increases the recognition rate and confirms that the standard DRM configurations are not optimal for V1-V2 sequence recognition.

As for recognition rate, tables TT-2-2 and TT-2-3 give acceptable and consistant results.

However, we note that the configurations of the table TT-2-2 allow the model to produce trajectories closer to natural speech. Table TT-2-2 contains configurations selected using formant patterns from vowels in context for each speaker, most of the time this table provides optimized initial and final configurations of the transition and therefore, the corresponding total acoustic distance decreases.

The recognition rate obtained and the quality of the model trajectories depend on the codebook quality and on the model capacity to span the speaker's acoustic space. The configurations used are always realistic, i.e physically reachable by a human vocal tract. This acoustic space is now described for each vowel.

For [i], [œ], [ø]: the third formant values are too low, they never reach the speaker's acoustic space, for [a], [o] and [u]: F3 is too high and never meets the speaker's space; furthermore for [a], F1 is too low and for [i] F2 too high. The model acoustic space matches that of the other vowels for each formant value.

Figure 1 compares the three cardinal vowels [i], [a] and [u] with the DRM model configurations used for speaker TS with the area function proposed by [3].

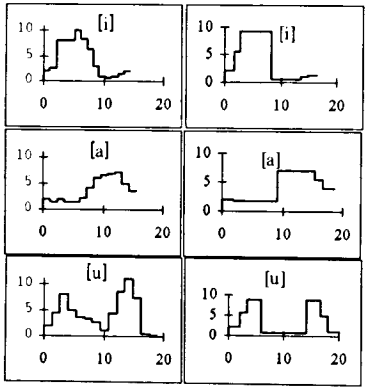


figure 1: Comparison of the three cardinal vowels [i], [a] and [u] with the DRM model configurations (rigth) used for speaker TS with the area function proposed by Majid [3](left). TheX-axis represents the distance from the glottis (cm), the Y-axis represents the regions' areas (cm^2).

Some V1-V2 transitions raise problems. For example [i]-[u] is well recognized when table TT-2-3 is used but the model formant trajectory badly matches the speaker's, the model cannot represent how the F2 and F3 formants cross, which is a characteristic of this transition. Transition [a]-[i] also has a good recognition rate and a poor model representation. For [i]-[y], not only is the third formant of [i] never reached, but also the F2 formant model value is not constant throughout the transition. Model transition with [e] or [ɛ] are acoustically close to the natural curve because these two vowels have good static representations and therefore the total distance decreases.

The transition [ɔ]-[a] always has a good recognition rate when table TT-2-2 and TT-2-3 are used for speakers TS and FB and is always well modelized for each speaker with table TT-2-2 (see figure 2). V1-V2 sequences with [a] also have a good recognition rate.

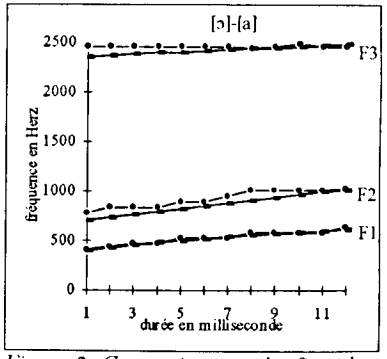


Figure 2: Comparison on the first three formants between the model trajectory ("-") and speaker's ("■ ") for [ɔ]-[a].

3.2 - Dissymetric model

If the symetric axis of the [i] configuration is shifted by 1 cm towards the glottis, the third formant better matches natural vowel space. However, this shift disturbs the [y] acoustic space. This vowel needs the symetric axis to be shifted towards the lips. This leads to change the symetric axis for each vowel and then to change the model characteristics continuously.

The results are not improved if the symetric axis shift is set identically for every vowel. So we keep the initial symetric tube. This shift of the symetric axis of the vocalic tube could be interesting to process a female voice since we know that the characteristics of the female vocal tract are different from those of the male vocal tract.

4 - CONCLUSION

The Distinctive Region Model study leads us to specify the capacities of this production system within the framework of vocal recognition. The constraints we imposed seem to be strong enough to avoid the one-to-many problem. Besides, we need to test different acoustic distances and acoustic parameters to measure the match between the model productions and the speaker's realisations.

Our study shows clearly that speaker adaptation is necessary and significantly improves the results. Moreover, context dependent configurations produce more accurate results.

Some speaker transitions are faithfully represented by the DRM model and have a good recognition rate. However, some particular cases remain to be examined more precisely.

The problem of the acoustic values not reached by the model is not resolved by a dissymetric model. The results are not improved by the use of different interpolation types or the desynchronisation of lip movement.

Generally speaking, our attempt to refine the recognition strategy doesnt improve the model performances. This model is very simple, it cannot respond precisely to any situation but it helps us to obtain suitable results for the vocalic transition recognition. The results obtained, on average 80% of the V1-V2 transition correctly identified, encourage us to carry on speech recognition with the articulatory models. Nevertheless it is advisable to use much more complex models able to take into account the whole articulatory phenomenon (recognition of the articulation mode and place for consonants).

ACKNOWLEDGEMENTS

We would like to thank Cecile Bianchi for her help in translating the article.

REFERENCES

[1] Candille L., H. Méloni, T. Spriet, R. Carré (1994), "Inversion de modèle articulatoire pour la reconnaissance de la parole: application à l'identification de diphones vocaliques",20e JEP, Trégastel, p: 485-490

[2] Carré R. & Mrayati M. (1992), "Distinctive Regions in acoustic tubes. Speech production modeling."Journal d'acoustique 5, 141-159.

[3] Majid R. (1986), "Modélistaion articulatoire du conduit vocal, exploration et exploitation. Fonction de macrosensibilité paramétriques et voyelles du français" Thèse Doc. Ing., INP Grenoble.

BIMODAL RECOGNITION OF ITALIAN PLOSIVES

P. Cosi, M. Dugatto, F. Ferrero, E. Magno Caldognetto and K. Vagges
Centro di Studio per le Ricerche di Fonetica (CNR)

ABSTRACT

A bimodal automatic speech recognition system, in which the speech signal is synchronously analyzed by an audio channel producing spectral-like parameters every 2 ms and by a visual channel computing lip and jaw kinematic parameters, is described and some results are given for various speaker independent phonetic recognition experiments regarding the Italian plosive class in different noisy conditions.

INTRODUCTION

Audio-visual automatic speech recognition (ASR) systems can be conceived with the aim of improving speech recognition performance, mostly in noisy conditions [1]. Various studies of human speech perception have demonstrated that visual information plays an important role in the process of speech understanding [2], and, in particular, "lip-reading" seems to be one of the most important secondary information sources [3]. Moreover, even if the auditory modality definitely represents the most important flow of information for speech perception, the visual channel allows subjects to better understand speech when background noise strongly corrupts the audio channel [4]. Mohamadi and Benoît [5] reported that vision is almost unnecessary in rather clean acoustic conditions (S/N > 0 dB), while it becomes essential when the noise highly degrades acoustic conditions (S/N ≤ 0 dB).

METHOD

The system being described takes advantage of jaw and lip reading capability, making use of a new system for automatic jaw and lips movement 3D analysis called ELITE [6], in conjunction with an auditory model of speech processing [7] which have shown great robustness in noisy condition [8].

The speech signal, acquired in synchrony with the articulatory data, is prefiltered and sampled at 16 KHz, and a joint synchrony/mean-rate auditory model of speech processing [7] is applied producing 80 spectral-like parameters at 500 Hz frame rate. In the experiments being described, spectral-like parameters and frame rate have been reduced to 40 and 250Hz respectively in order to speeding up the system training time. Input stimuli are segmented by SLAM, a recently developed semi-automatic segmentation and labeling tool [9] working on auditory model parameters.

The visual part of the system has adopted ELITE which is a fully automatic movement analyzer for 3D kinematic data acquisition. This system ensures a high accuracy and minimum discomfort to the subject. In fact, only small, non obtrusive, passive markers of 2mm of diameter, realized by reflective paper, are attached onto the speaking subject's face. The subjects are placed in the field of view of two CCD TV cameras at 1.5 meters from them. These cameras light up the markers by an infrared stroboscope, not visible in order to avoid any disturbance to the subject. ELITE is characterized by a two level architecture. The first level includes an interface to the environment and a fast processor for shape recognition (FPSR). The outputs of the TV cameras are sent at a frame rate of 100 Hz to the FPSR which provides for markers recognition based on a cross-correlation algorithm implemented in real-time by a pipe-lined parallel hardware. This algorithm allows the use of the system also in adverse lighting conditions, being able to discriminate between markers and reflexes of different shapes although brighter. Furthermore, since for each marker several pixels are recognized, the cross-correlation algorithm allows the computation of the weighted center of mass increasing the accuracy of the system up to 0.1mm on 28cm of field of view. The coordinates of the recognized markers are sent to the second level which is constituted by a general purpose personal computer. This level provides for 3D coordinate reconstruction, starting from the 2D perspective projections, by means of a stereophotogrammetric procedure which allows a free positioning of the TV cameras. The 3D data coordinates are then used to evaluate the parameters described hereinafter.

Finally both audio and visual parameters, in a single or joint fashion, are used to train, by means of the Back Propagation for Sequences (BPS) [10] algorithm, an artificial Recurrent Neural Network (RNN) to recognize the input stimuli.

A block diagram of the overall system is described in Figure 1 where both the audio and the visual channel are shown together with the RNN utilized in the recognition phase.

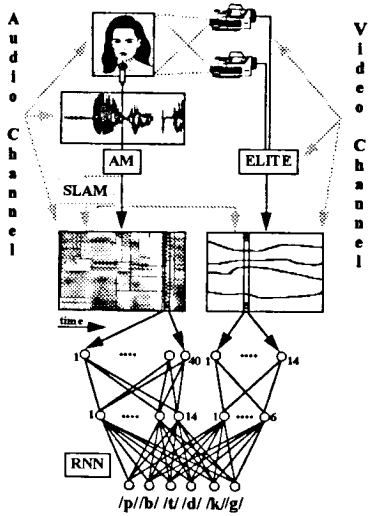


Figure 1. Structure of the bimodal recognition system.

EXPERIMENT

The input data consist of disyllabic symmetric /'VCV/ nonsense words, where C=/p,t,k,b,d,g/ and V=/a,i,u/, uttered by 10 male speakers. All the subjects were northern Italian university students, aged between 19 and 22, and were paid volunteers. They repeated five times, in random order, each of the selected non-sense words. The speaker comfortably sits on a chair, with a microphone in front of him, and utters the experimental paradigm words, under request of the operator. Three reference points and five target points on the face of the subject have been considered. As illustrated in Figure 2, these points are the nose (n.1), the middle edge of the upper lip (n.2), the middle edge of the lower lip (n.5), the corners of the mouth (n.3 and n.4), the jaw (n.6), and the lobe of the ears (n.7 and n. 8).

In this study, the movements of the markers placed on the central points of the vermilion border of the upper lip (marker 2), and lower lip (marker 5), together with the movements of the marker placed on the edges of the mouth (markers 3, 4), were analyzed, while the markers placed on the tip of the nose (marker 1), and on the lobe of the ears (markers 7, 8), served only as reference points. In fact, in order to eliminate the effects of the head movement, the opening and closing gestures of the upper and lower lip movements were calculated as the distance of the markers 2 and 5 placed on the lips, from the plane depicted in Figure 2 and defined by the line passing from the markers 7 and 8, placed on the ear lobes, and marker 1, placed on the tip of the nose. Similar distances with a plane perpendicular to the above one serve as a measure of upper and lower lip protrusion. A total of 14 values, defined as the difference between various markers or between markers and reference planes, plus the correspondent instantaneous velocity, obtained by numerical differentiation, constitute the articulatory vector which has been used together with the acoustic vector in order to represent the target stimuli. The articulatory parameters were, besides the upper and lower lip opening and closing movements, and the upper and lower lip protrusion, the lip opening height calculated as the distance between markers 2 and 5, the lip opening width, calculated as the distance between markers 3 and 4, and the jaw opening measured between the markers placed on the jaw and on the tip of the nose.

As an example of the articulatory parameters, Figure 3 shows the vertical displacement and the instantaneous velocity of the marker placed on the lower lip (n. 5) associated with the sequence /'apa/.

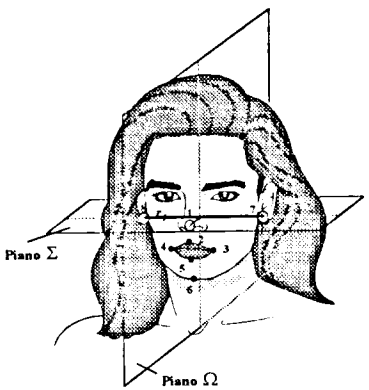


Figure 2. Position of the reflecting markers and of the reference plane. Identification numbers are indicated next to their corresponding markers. Marker dimension in the figure does not correspond to the real dimension (2mm) but is exaggerated for visualization purpose.

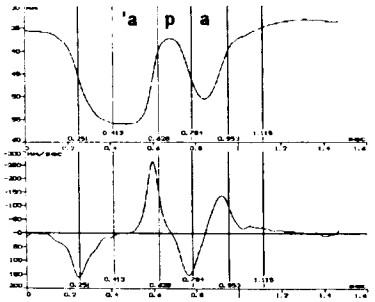


Figure 3. Time evolution of displacement and velocity of the marker placed on the lower lip (n.5), associated with the sequence /'apa/.

In a previous work [11], regarding the same task, but in a Speaker Dependent (SD) environment, comparing results obtained in the three considered situations:

a) only the audio channel is active;
b) only the visual channel is active;
c) both audio and visual channel are simultaneously active,

the usefulness of visual parameters for improve speech recognition performance was successfully demonstrated. In this work, regarding a Speaker Independent (SI) environment, only the first and third situations in which only the audio channel or both audio and visual channels are simultaneously active were considered.

The network architecture which has been considered for the recognition was a fully connected recurrent feed-forward BP network with dynamic nodes positioned only in the hidden layer. The learning strategy was based on BPS algorithm [10], and only two supervision frames were chosen in order to speeding up the training time. The first one, focused on articulatory parameters, was positioned in the middle frame of the target plosive, the 'closure' zone, while the second, focused on acoustic parameters, was positioned in the penultimate frame, the 'burst' zone. A 20 ms delay, corresponding to 5 frames, was used for the hidden layer dynamic neurons. A 54(40+14)input * 20(14+6)hidden * (6)output RNN, as illustrated in Figure 1, was considered. Not all the connections were allowed from the input and the hidden layer, but only those concerning the two different modalities, which were thus maintained disjoint. Various parameter reduction schemes and various network structure alternatives were exploited but those described above represent the best choice in terms of learning speed and recognition performance.

RESULTS

Two different experimental setting were considered in which, among the 10 speakers, 8 speakers were randomly picked up in order to form the learning set while the remaining two were considered as the test set. The results for these two cases are illustrated in Table 1.

Table 1. SI correct recognition rate in two experimental settings with 8 speaker for learning and 2 for testing.

	E1	E2
Speaker 1	95.6	87.8
Speaker 2	72.2	66.7
Mean	83.9	77.8

After having observed that a particular speaker had a vary bad acquired audio signal, a third experiment was organized thus considering only 7 speakers for the learning set and two for the test set. Results regarding this case are summarized in Table 2.

In Table 3 the Speaker-Pooled (SP) mean correct recognition performance for all the three experimental settings is illustrated. In this case each speaker forming the learning set was also individually tested.

Table 2. SI correct recognition rate with the 9 speaker set (see text).

	E3
Speaker 1	95.6
Speaker 2	84.4
Mean	83.9

Table 3. SI mean correct recognition rate for the Speaker-Pooled (SP) case.

	E1	E2	E3
Mean	78.5	74.8	83.3

In order to test the power of the bimodal approach all the three experiments were repeated eliminating visual information thus retaining only the audio channel input. The 40 input * 14 hidden * 6 output RNN utilized in this case is exactly the audio subnet of the global net utilized in the bimodal environment as indicated in Fig. 1.

Table 4. SI mean correct recognition rate with only Audio information.

	E1	E2	E3
Mean	68.9	58.3	65.0

CONCLUSIONS

As indicated by a direct inspection of Tables 1-4, recognition performance significantly improves when both audio and visual channels are active. Looking at Table 3 referring to the speaker-pooled results a good generalization power can be associated with the chosen RNN given that SI results were surprisingly better than SP results.

REFERENCES

[1] P.L. Silsbee and A.C. Bovik (1993), "Medium-Vocabulary Audio-Visual Speech Recognition", *Proc. NATO ASI, New Advances and Trends in Speech Recognition and Coding*, pp. 13-16.

[2] D.W. Massaro (1987), "*Speech Perception by Ear and Eye: a Paradigm for Psychological Inquiry*", Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[3] B. Dodd & R. Campbell, Eds., (1987), "*Hearing by Eye: The Psychology of Lip-Reading*", Lawrence Erlbaum Associates, Hillsdale, New Jersey.

[4] A. MacLeod and Q. Summerfield (1987), "Quantifying the contribution of vision to speech perception in noise", *British Journal of Audiology*, 21 pp. 131-141.

[5] C. Benoit (1992), "Bimodal Aspects of Speech Communication", personal communication.

[6] G. Ferrigno and A. Pedotti (1985), "ELITE: A Digital Dedicated Hardware System for Movement Analysis via Real-Time TV Signal Processing", *IEEE Transactions on Biomed. Eng.*, BME-32, pp. 943-950.

[7] S. Seneff (1988), "A joint synchrony/mean rate model of auditory speech processing", *Journal of Phonetics*, 16, 1988, pp. 55 76.

[8] P. Cosi(1992), "Ear Modelling for Speech Analysis and Recognition" (1992). In M. Cooke, S. Beet and M. Crawford (Eds.), *Visual Representation of Speech Signals*. John Wiley & Sons, pp.205-212.

[9] P. Cosi (1993), "SLAM: Segmentation and Labelling Automatic Module", *Proc. Eurospeech-93, Berlin*, 21-23 September, 1993, pp. 665-668.

[10] M. Gori, Y. Bengio and R. De Mori (1989), "BPS: A Learning Algorithm for Capturing the Dynamical Nature of Speech", *Proc. IEEE IJCNN89*, Washington, June 18 22, 1989, Vol. II, pp. 417 432.

[11] P. Cosi, E. Magno Caldognetto, K. Vagges, G.A. Mian, and M. Contolini (1994), "Bimodal Recognition Experiments with Recurrent Neural Networks", *Proceedings of IEEE ICASSP-94*, Adelaide, Australia, 19-22 April, 1994, paper 20.8.

EXPLOITING THE AUDITORY INDUCTION EFFECT IN ROBUST SPEECH RECOGNITION

Malcolm Crawford, Martin Cooke and Phil Green
The Institute for Language Speech and Hearing, The University of Sheffield
West Court, 2 Mappin Street, Sheffield S1 1DT
{m.crawford, m.cooke, p.green}@dcs.shef.ac.uk
http://www.dcs.shef.ac.uk/research/ilash/

ABSTRACT

This paper builds on previous work on recognition of speech in noise to incorporate a model of the constraints imposed by rules governing auditory induction, implemented as a refinement of the distance metric used in a Kohonen network. We show that use of this constraint improves recognition accuracy, and points to a new understanding of some aspects of speech perception.

INTRODUCTION

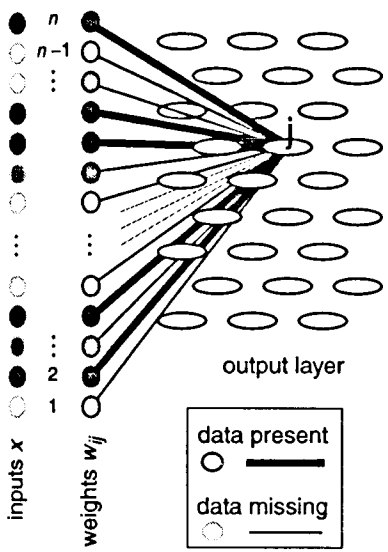
It is a matter of everyday experience that speech perception is possible in "adverse" conditions. Bregman [1] has coined the term *auditory scene analysis* to refer to listeners' ability to separate out individual sound sources in a mixture by grouping together those components of a signal which share characteristics in common (e.g. harmonicity). Recent evidence [2] has suggested, however, that in noisy environments it is not possible for the auditory system to fully recover all the evidence for, say, a speech signal. It is often the case that an intrusive noise source will completely dominate a particular time-frequency region: in these regions there are no components which can reliably be ascribed to the speech signal, so its representation will necessarily be incomplete. Few theories of speech perception, however, consider the effects that extraneous signals have on the auditory representation of speech signals, or account for listeners' abilities to induce the percept of the continuity of a speech signal through noise.

Those parts of the spectrum which cannot be attributed to the speech signal are nevertheless information-bearing. They place an upper bound on the amount of energy that the speech signal could have had. If, for example, evidence is only available for low frequencies (in the F1 region) and this suggests initially that an /i/ might be present, we would minimally expect also to find concentrations of spectral energy in the 2200-3300Hz region, corresponding to F{2,3,4} of the vowel. If this is not observed, it constitutes evidence counter to the initial /i/ hypothesis. This area has been formally investigated in studies of auditory induction: Warren and his colleagues [3] have found that the auditory system is indeed subject to such constraints. In their experiments, part of a stimulus signal is excised and replaced with a noise burst. If and only if the noise is sufficiently loud that it would have masked the original sound, had it been present, the auditory system induces the percept of the original, leading to continuity.

We now show how the auditory induction constraint can be incorporated into a recognition architecture adapted to handle missing data in a bottom-up fashion.

RECOGNITION FROM PARTIAL DATA

We have developed [4] a model of speech recognition, based on a modified Kohonen self-organising network [5], whose performance degrades gracefully as an increasing proportion of the input representation of speech is deleted (simulating the effect of increasing noise intrusions) during training and recognition. During the recognition phase, an input vector is compared with the weight vector



Normal distance metric [5]	$\sum_{i \in present} (w_{ij} - x_i)^2$
Distance metric for incomplete vectors [6]	$\sum_{i \in present} (w_{ij} - x_i)^2$
Distance metric with auditory induction constraint	$\sum_{i \in present} (w_{ij} - x_i)^2 + \sum_{i \notin present} max(0, w_{ij} - x_i)^2$

Figure 1. The input vector is compared with net weight vectors: the winning unit is chosen as that whose weight vector most closely matches the input vector, as determined by a distance metric.

associated with every node in the net. In the general case, all components of the input vector contribute to the distance measure. In the version modified for incomplete input [6] only those components present in the input vector contribute to the score.

As discussed previously, studies of perceived auditory continuity suggest that the full spectral profile places additional top-down constraints on the recognition process. Specifically, "auditory induction" operates only if sufficient energy exists in the occluded regions to account for the induced pattern. This is expressed computationally by the addition of a second factor in the distance metric which adds a penalty for any "absent" components whose level is below that in the net vector. This is illustrated in figures 1 and 2.

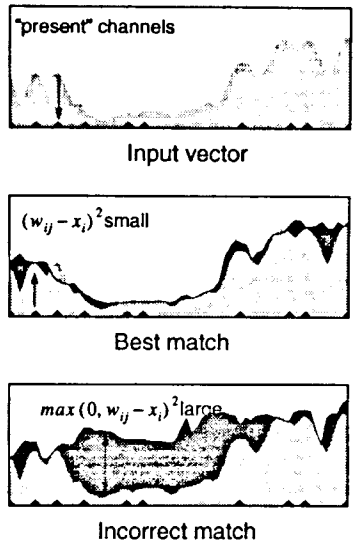


Figure 2. Illustration of the auditory induction constraint: the "best match" panel shows a correct match between the input vector (light grey) and the net weight vector (dark grey). The distance measure is determined primarily by standard metric. The "incorrect match" panel shows a match between the same input vector and a net weight vector which "expects" more energy than is present in the signal.

EXPERIMENTS

A net of dimensions 19x13 was used. The input representation was produced by a 64-channel gammatone filterbank, with channel centre-frequencies equally spaced on an ERB-rate scale between 200 and 5500 Hz, and the output of each filter

processed by a model of inner hair cell transduction, smoothed over a 10 ms window.

Training and test data were generated from utterances produced by a single male Japanese speaker from the ATR large-scale speech database [7]. The nets were trained and calibrated (i.e. one of 27 phone labels was attached to each output node) using a training set, and recognition performance measured in terms of recognition accuracy — the percentage of labels in the test set correctly identified.

Recognition performance was investigated in two series each of 10 different conditions, in which during recognition input vector components were deleted at random with a probability which varied in from 0.0 (no deletion) to 0.9 (90% deletion). In the first series the distance metric for incomplete vectors was used, in the second the metric (AIC) with auditory induction constraint was employed.

Results of these experiments are shown in figure 3. They show a clear benefit of using the AIC metric as the probability of deletion increases (using a model of auditory scene analysis the probability of deletion is likely to be around 85%).

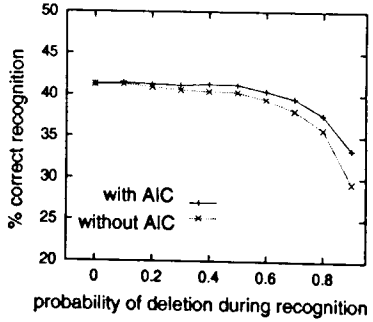


Figure 3. Recognition accuracy vs. probability of deletion for 2 recognition algorithms. The "without AIC" curve employs the distance measure for incomplete vectors, whilst "with AIC" includes modifications suggested by the continuity effect as described in the text.

IMPLICATIONS FOR MODELS OF SPEECH PERCEPTION AND FUTURE RESEARCH

These results point to a model of speech perception in which information grouped by auditory scene analysis triggers matches against stored speech schemas, which may then be verified through application of the auditory induction constraint.

This further suggests that the representations the auditory system uses to "categorise", and indeed learn about, speech sounds may be rather different to those with which the traditional phonetician or linguist familiar. If we restrict ourselves to consideration of spectrum-like representations (there is no reason why other representations such as onset, offset and frequency-transition maps should not play a part in the coding of speech), this is especially true in low-frequency regions of speech, where the first formant, and in some cases the second, and even third formants, are resolved into harmonics. When this has been addressed (cf. [8]), it has generally posed a problem for theories of speech perception, as well as automatic speech recognition systems.

A representation which includes harmonics will clearly be (even) more variable than one in which the formants are coherent, due to natural changes in fundamental frequency during the course of phonation. A model of perception based on partial matching suggests that we should regard grouped auditory primitives as indicating those frequencies at which the spectrum should be sampled, rather than as defining a pattern to be matched. This side-steps the problems associated with smoothing reconstructed spectral profiles on the basis of extracted peaks using a process of interpolation. Furthermore, a partial matching scheme might account for the centre of gravity effect, or large-scale spectral integration (cf. [8]). It gives a plausible mechanism by which the auditory system would take into account gross spectral shapes rather than individual formant frequencies. Consider for example the weights in nets *trained* (cf. [4]) using complete data vectors and data vectors with 85% of the components randomly deleted, shown in figure 4. It can be clearly seen that the deletions have little effect on the representation of the vowel spectra.

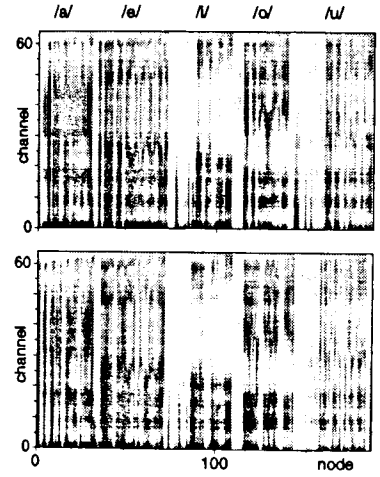


Figure 4. Spectrogram-like plots of net weight vectors (1 frame = 1 node) sorted by label (within-label order is insignificant) for nets trained using (top) complete data vectors and (bottom) data vectors with 85% of the components randomly deleted.

The model presented here may also explain aspects of the Lombard effect (cf. [9]): since harmonicity is a powerful grouping cue, and in noisy conditions likely to be more robust than common amplitude modulation, a goal of the elevation of pitch may be to make the voice more readily separated. Furthermore, studies of auditory induction [3] also show that when one sound is heard in the presence of another, the perceived loudness of the first sound is less than it would be were it heard in isolation. In other words the auditory system employs some sort of disjoint allocation of energy between the two sounds. The purpose of changes in spectral tilt observed in Lombard speech may be to raise the perceived level of high frequency regions of the spectrum to counteract the fact that otherwise, after disjoint allocation, the perceived levels would be below those expected for a particular speech sound.

Finally, the model suggests a novel methodology for investigating phonetic cues, and cue trading: it makes it possible to investigate computationally the question "can recognition be achieved without information in this region".

REFERENCES

[1] A.S. Bregman (1990), *Auditory Scene Analysis*, MIT Press.

[2] G.J. Brown & M.P. Cooke (1994), "Computational auditory scene analysis", *Computer Speech & Language*, **8**, 297-336.

[3] R.M. Warren, J.A. Bashford Jr., E.W. Healy & B.S. Brubaker (1994), "Auditory induction: Reciprocal changes in alternating sounds", *Perception & Psychophysics*, **55** (3), 313-322.

[4] M.P. Cooke, P.D. Green & M.D. Crawford (1994), "Handling missing data in speech recognition", *International Conference on Speech and Language Processing*, Yokohama.

[5] T.E. Kohonen (1984), *Self-Organisation and Associative Memory*, Springer.

[6] T. Samad & S.A. Harp (1992), "Self-organisation with partial data," *Network*, **3**, 205-212.

[7] A. Kurematsu *et al.* (1990), "ATR Japanese speech database as a tool of speech recognition and synthesis", *Speech Communication*, **9**, 357-363.

[8] A. Bladon, (1986), "Phonetics for hearers", in: *Language for Hearers*, ed: G. McGregor, Pergamon Press.

[9] J-c. Junqua (1992), "The variability of speech produced in noise", *Proc. ESCA Workshop on Speech Processing in Adverse Conditions*, Cannes, 187-190.

ACKNOWLEDGEMENTS

This work was supported by a study visit grant to ATR, Kyoto to Malcolm Crawford, and by SERC Image Interpretation Initiative Research Grant GR/H53174. Kohonen net simulations adapted the public domain SOM_PAK code (Kohonen, Kangas and Laaksonen, 1992).

A BOTTOM UP HYBRID METHOD FOR ISOLATED WORDS RECOGNITION

Olivier Delemar & Harouna Kabré
Institut de la Communication Parlée, Grenoble, France

ABSTRACT

In this article we present a new hybrid method for speech recognition which uses expert knowledge to control Hidden Markov Models in a purely bottom-up way. The phonetic knowledge used by the hybrid model to perform this control is embedded in the standard hidden Markov models by the way of an "expert matrix" which expresses whether a given broad phonetic label may or may not, be detected by the expert system while the model is in a given state.

INTRODUCTION

Many recent works in the field of speech recognition tend to combine different methods of acoustic-phonetic decoding, in order to take advantage of their particularities. Among the different hybrid methods, there are those which use the discrimination power of neural networks with the time alignment capacities of Hidden Markov Models [1] [2] [3] and those which combine rule based systems with HMM [4] or neural networks [5].

Although the Markov modeling is one of the best speech recognition methods, it still has great difficulty coping with explicit phonetic knowledge. A usual solution to this problem is to constrain HMM to assign one state either for one particular acoustic phase of the speech unit [6] [7] [8], or for one articulatory configuration of the vocal track [9]. The underlying principle of these methods is that they force the matching of the underlying phonetic structure - which may be described by a human expert - by the state transition graph of the Markovian process.

This paper presents a hybrid recognition method which uses expert rules to add automatically phonetic knowledge to the set of standard HMM parameters during the training step and to control the bottom-up recognition process according to that knowledge.

The rest of the paper is organized as follows: the first section will present the standard HMM principle, then the hybrid model will be described with an example showing how it controls the recognition algorithm. The last section will give some results and will discuss the enhancements and limitations of this method.

HIDDEN MARKOV MODELING

HMMs are finite states automatons which model sequences of quasi-stationary phases [10]. They are formed by a fixed number of states linked to others by arcs. Each arc has an associated transition probability - possibly null - while each state is associated with an emission probability density function (pdf). An N-states Markov model is then defined entirely by it's A-matrix of transition probabilities a_{ij} and it's set of emission pdf's $b_i(.)$ called the B-matrix.

Speech recognition with HMM consists in evaluating each model probability, given a vector of acoustic features. This may be performed by the Viterbi algorithm [11] which, in addition, finds the best path along the Markov chain in order to maximize this probability. During this process, the choice of a transition from state i to state j, given the feature vector O_t, depends on the value $a_{ij}*b(O_t)$ which may be seen as the "cost" of the transition from state i to state j, weighted by the "distance" between the j^{th} state's inner representation of an acoustic configuration and the observed acoustic feature O_t. Thus the Viterbi algorithm performs nothing other than a time alignment procedure. Figure 1 shows such an alignment for the French word "ouvre" (/uvR@/).

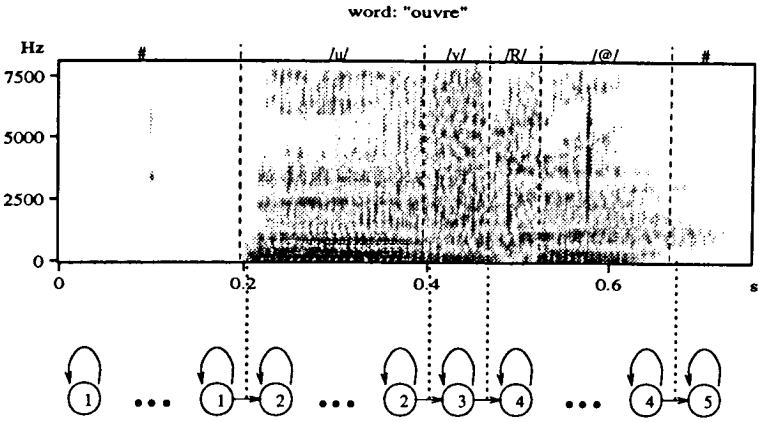


Figure 1: Time alignment of the word model "ouvre" with an occurrence of this word: the transitions from one state to the next one are quite exactly synchronized with the manually determined phonemes' frontiers.

The efficiency of the acoustic-decoding with HMM relies on the optimization of the models' parameters. The forward-backward algorithm [12] which is widely used for this purpose is an Expectation-Maximization procedure which iteratively re-estimates the transition and emission probability, given a training set of acoustic data. It should be noticed that, as far as HMMs are probabilistic models, the number and the quality of the training samples condition the representativity and the generalization capacity of the models. Furthermore, short acoustic events like the stop-consonants' burst are poorly modeled because of their reduced number of representative feature vectors.

THE HYBRID MODEL

The hybrid model presented here tries to satisfy three constraints: introducing phonological knowledge expressed by an expert system into the HMM's decision procedure, keeping the automatic aspect of the model's training phase and maintaining the bottom-up aspect of the acoustic-phonetic decoding by HMM which allows a real-time implementation for speech recognition. Taking into account the speech segmentation generated by the Viterbi algorithm, the training phase of the hybrid model will create a so called "expert matrix" E with as many rows as the Markovian model has states and as many columns as there are broad phonetic classes predicted by the expert system.

In our experiments, the expert system is a set of deterministic networks [13] finding occurrences of voiceless fricatives and stop consonants by applying fuzzy thresholds to the zero-crossing rate, the power and its first and second order derivative. Thus, after a standard model training, a time alignment is achieved for each training sample and compared to the broad phonetic labels in order to evaluate the "plausibility" that this label will be predicted by the expert system while the i^{th} state of the model is being visited.

During the recognition phase, both Markov models and the expert system are run simultaneously. Each time a label is generated by a deterministic network, the expert matrices are parsed to determine the states of each models which may be visited at this time. All other states are weighted so that any state sequence containing these states is forbidden. The constraints applied to HMM are then dynamic, and the hybrid model is still bottom-up. They are also time synchronous and the expert knowledge is then taken into account by the time alignment process.

RESULTS AND DISCUSSION

The hybrid model has been tested on a subset of the French database BDSON. This corpus is formed by 161 mono or bi-syllabic words, each pronounced once

by 5 male and 5 female speakers. The speech signal is sampled at 16kHz and analyzed every 10ms by a Perceptual Linear Predictive Coding algorithm [14] to produce the feature vectors. Both standard and hybrid models are trained with 6 of the 10 utterances and the other 4 are used as the test set.

For this corpus, the standard HMMs give a 38% recognition rate (49% if we consider the first two candidates). By correcting 16% of the confused words, the hybrid model increases this rate to 43% (56% for the first two candidates).

word "autre" is not penalized because the best states sequence implies that the appropriate state is visited when the voiceless plosive is detected and this model becomes the most likely one.

As a result of the time synchronous constraints imposed on the models' states sequences, the hybrid model demonstrates on the ability to closely model the phonetic structure of words, even with a small number of training utterances. This is due to the fact that the a priori knowledge-based expert system is not dependent on the quantity of training data. Furthermore, the principle

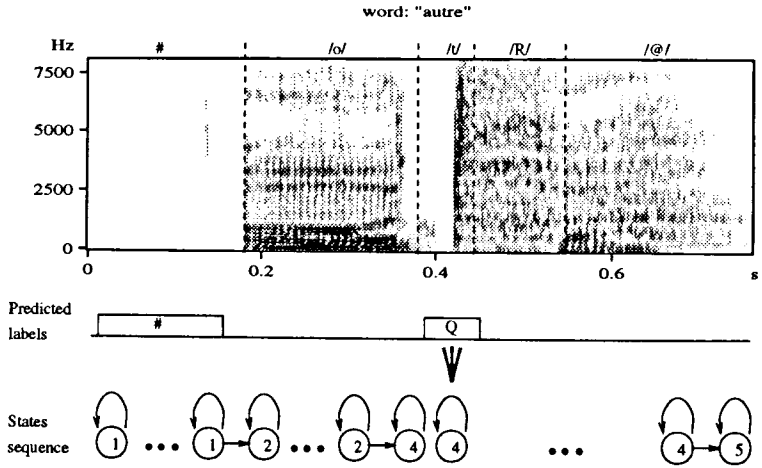


Figure 2: How the expert system controls the Markovian model: the path followed by the Viterbi algorithm along model's states is penalized when the "Q" label is predicted by the expert system.

An example of wrong-model correction is shown in figure 2. As a result of differences in pronounciation between the speaker and those used to train the models, the uttered word "autre" (/otR@/) happens to be more likely emitted by the model of the word "ouvre". On the other hand, acoustic events like the short silence followed by a noised burst are clearly detected by the expert system which then produces the label Q, indicating a voiceless plosive. As long as there is no state -according to the E matrix- which may be visited while this phonetic label occurs, all states' probability are weighted down, resulting in a reduction in the model's final probability. By contrast, the model of the

of time rendez-vous imposed on HMMs' states sequences by the external system may be extended to handle other types of information.

The principle of this method is that it uses the external information whenever it is available, this means that the expert system has to be as robust as possible. In fact, some unexpected detections made by the expert system cause the correct model to be penalized. Thus further works will tend to optimize the performance of the expert system in order to close approach the results obtained in a preliminary experiment, using manualy given labels obtained from a human expert [15].

CONCLUSION

We have described a new hybrid approach to speech recognition using HMMs and a rule-based expert system. Phonological knowledge as expressed by the expert system is embedded in the models by the way of the E matrix. This knowledge is then used during a purely bottom-up recognition process, to constrain the states sequence and avoid forbidden states. The results are encouraging but the rules have to be optimized in order to produce solely robust information.

REFERENCES

[1] Franzini M. A., Witbrock M. J. and Lee K-F (1989). "A Connectionist Approach to Continuous Speech Recognition", *ICASSP 89*, vol. 1, pp. 425-428.

[2] Bourlard H. and Wellekens C. J. (1990) "Links between Markov Models and Multilayer Perceptrons", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 1167-1178.

[3] Renals S., Morgan N., Bourlard H., Cohen M. and Franco H. (1994). "Connectionist Probability Estimators in HMM Speech Recognition", *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, pp 161-174.

[4] Haton J. P., Carbonell N., Fohr D., Mari J. F. and Kriouille A. (1987) "Interaction between Stochastic Modeling and Knowledge-Based Techniques in Acoustic-Phonetic Decoding of Speech", *Proc. ICASSP*, vol. 2, pp. 868-871.

[5] Nocera P. and Bulot R. (1994) "Utilisation d'une méthode mixte : expertise et réseaux de neurones pour la reconnaissance des occlusives du français", *Reconnaissance automatique de la parole*, GDR-PRC Communication Homme-Machine.

[6] Deng L., Lenning M. and Mermelstein P. (1990) "Modeling Microsegments of Stop Consonants in a Hidden Markov Model Based Word Recognizer", *JASA 87(6)*, pp. 2738-2747.

[7] Farhat A., Pérennou G. and André-Obrecht R. (1993) "A Segmental Approach versus a Centiseconde one for Automatic Phonetic Time-Alignment", *EUROSPEECH 93*, vol. 1, pp. 657-660.

[8] Jouvet D., Barthova K. and Mouné J. (1991) "On the Modelization of Allophones in an HMM-Based Speech Recognition System", *EUROSPEECH 91*, vol. 2, pp 923-926.

[9] Deng L. and Erler K. (1992) "Structural Design of Hidden Markov Model Speech Recognizer using Multivalued Phonetic Features: Comparison with Segmental Speech Units", *JASA 92(6)*, pp. 3058-3067.

[10] Rabiner L. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE*, vol. 77, pp. 257-285.

[11] Viterbi A. J. (1967). "Error Bounds for Convolution al Codes and an Asymptotically Optimal Decoding Algorithm", *IEEE Trans. Informat. Theory*, vol. IT-13, pp. 260-269.

[12] Baum L. E. (1972). "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes", *Inequalities*, vol. 3, pp. 1-8.

[13] Kabré H. (1991) "Décodage acoustico-phonétique multilingue : système à base de connaissances et étiquetage automatique de corpus de parole", *PhD Thesis*, Université Paul Sabatier, Toulouse.

[14] Hermansky H. (1990) "Perceptual Linear Predictive Coding of Speech", *JASA 87(4)*, pp. 1738-1752.

[15] Delemar O. (1994) "Reconnaissance de mots enchaînés par une méthode hybride : réseau markovien et base de règles", *Proc. 20èmes Journées d'Etude sur la Parole*, pp. 497-500.

PARALLEL DISTRIBUTED PROCESSES FOR SPEAKER-INDEPENDENT ACOUSTIC-PHONETIC DECODING

A. Ghio and M. Rossi
Institut de phonétique d'Aix-en-Provence
Laboratoire "Parole et Langage" URA 261, CNRS
29, Av.R.Schuman, 13621 Aix-en-Provence, FRANCE

ABSTRACT

We aim at examining to what extent a knowledge-based model can recognise segmental structure without feedback from semantic information and without stochastic modelling. The system proposed is inspired by some features of human cognitive processing: the speech signal activates parallel distributed processes of decoding. A supervisor module takes the final decision after the access to a dictionary and a top-down verification.

GENERAL PRESENTATION OF THE SYSTEM

The field of this study is automatic speech recognition and concerns more precisely speaker-independent acoustic-phonetic decoding. We aim at examining to what extent a knowledge-based model can recognise segmental structure without feedback from semantic information and without stochastic modelling.

The system proposed is inspired, in a functional way, by some features of human cognitive processing [1, 2]. The sequence of operations can be characterised as data driven. The speech signal first arrives at the low level analysis demons which then activate parallel distributed processes of decoding (Fig.1). The modules of this multi-analysis and multi-expert system are conceptually different. They consequently do not give the same output. Their results, then, are sent to the cognitive demons, who act upon them using high level information (phonological rules, access to a dictionary...). Finally, after a top-down verification, a decision process selects the alternative that has the strongest evidence.

First of all, we present the different modules of the bottom-up decoding i.e. the automatic segmentation, the global and the analytic recognition. Secondly, we develop the main ideas used in the high levels processes, especially in the access to a dictionary and the supervisor. Partial results are presented in a third part, just before a conclusion.

THE DIFFERENT MODULES OF THE BOTTOM-UP DECODING

The bottom-up decoding is composed by different parallel distributed processes.

The automatic segmentation

According to the general outlines of the Level Building procedure [3], an automatic segmentation module SAPHO (Segmentation by Acoustic-PHOnetic knowledge) supplies a hierarchical set of acoustic properties and segments, and phonetic properties and segments which fit the phonetic parsing of the acoustic wave [4]. It is not an unguided method, which are generally based on an instability function. In the SAPHO algorithm, energy + zero cross ratio parameter + some spectral features permit the location and the rough identification of segments. Only the nature of the segment authorises a posteriori the precise location of the boundaries.

The output is the labelling of temporal frames in macro-classes (vowel, stop, fricative, vocalic consonants, silence...) which permit an oriented analysis.

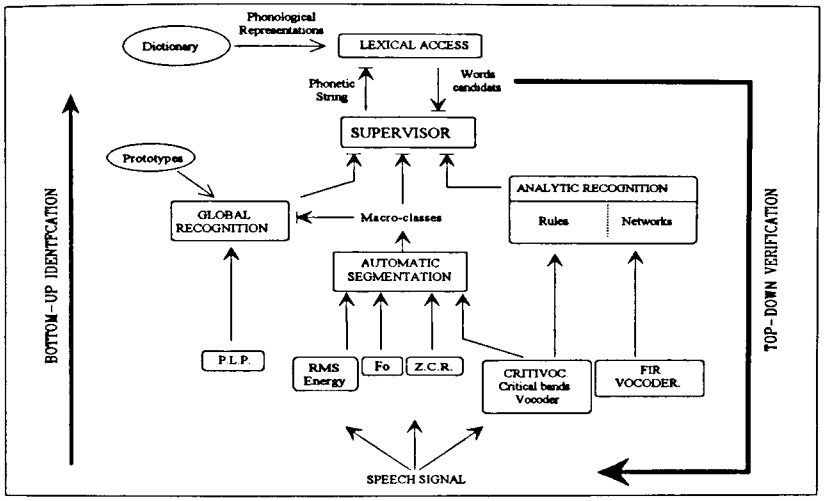


Figure 1 : Functional diagram of SYMULDEPHO

The global recognition

The global recognition module [5] is based on metric methods. Decoding units are CV groups. First, the feature extraction is done by a Perceptually based Linear Prediction analysis [6, 7]. Then, the system uses a Data Time Warping algorithm in order to compare stimuli to references. The output is a list of classified CV candidates. On account of the variability of speech, the system does not keep only the best. The analysis of cues relative to the ten best candidates allows the construction of the best prototype using a vote procedure (ex: if among the 10 best consonant candidates, 9 are voiced, 8 acute, 1 compact, 0 continuant, 1 nasal, 1 vocalic, the module propose [d] as solution)

The analytic recognition

The first analytic recognition module uses networks which are oriented graphs with state transitions. The differences with the Markov chains are the lack of probability and the fact that they work on the paradigmatic axis and not the syntagmatic one. They are supposed to model the allophones of vowels and not the abstract units. Each network is specialised for the recognition of one vowel. All are activated at each temporal frame. If a path is found along a network, an output appears at the end. The result is a table containing the allophone candidates. The analysis of the temporal distribution of phonemes leads to strong hypotheses on the vowel identification and its context.

The second analytic module is based on phonetic rules. Acoustic cues, different from the previous ones, are extracted every temporal frame (centi-second) by modelling some psycho-acoustic phenomena (weighted sound level, critical bands). The analysis of phonetic features allows the system to identify the phonemes using rules. A temporal tracking, which takes into account the contextual variability of segments, provides information by analysing acoustic transitions of coarticulation in triphones. For example, the sequence [...eeœœaaaaaaa...] is typically the result of the decoding of /a/ in the left context of /k, g/.

THE HIGH-LEVEL MODULES

Each parallel distributed process of decoding provides a set of phonetic units which are sent to the high-level modules.

The supervisor

The supervisor-process is not finished and we would like to improve the decision methods. Time being, the different results of the bottom-up decoding are received by the supervisor which then makes up a list of possible phonetic strings.

In the case of isolated words, the access to a dictionary allows the supervisor to classify the phonetic strings by associating them to words-candidates (cf. next section). The top-down verification, which will be described in a next paper, authorises the selection of the alternative that has the strongest evidence.

The access to a dictionary

Some methods of automatic spelling-correction compute a distance between a reference-word and a test-word; it relies on a series of operations that model errors of insertion, deletion and substitution [8]. It is possible to realise these operations using dynamic programming [9]. The module of lexical access is inspired by this method.

In our case, distance is not computed between graphemes but between the decoded phonetic string proposed by the supervisor and the phonetic representations of words stored in a dictionary (Fig.2). The dynamic programming is efficient to integrate in a single algorithm all the phenomena of insertion, deletion and substitution which appears in the bottom-up decoding.

The comparison requires the computation of a local distance between the sub-units of the strings (Fig.2). Whereas in the case of orthographic strings, the local distance is basic (0 if graphemes are the same, 1 if they are different), the case of phonetic strings requires a more sophisticated measure. Actually, the difference between /i/ and /e/ is less important than the confusion of /i/ with /p/. This is the reason why we have introduced a matrix of cost-confusion, which indicates the difference between each phoneme. It also authorises the non-precise definition of a phoneme in the stimulus string. For example, on Figure 2, the 5th unit of the stimulus has been decoded as 'liquid' which is the macro-class of /l/ and /r/.

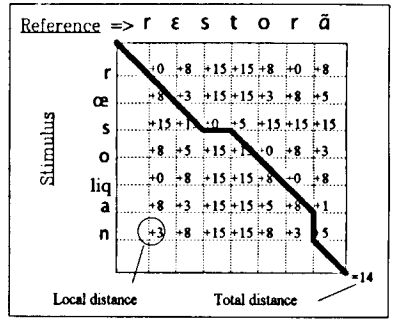


Figure 2 : Calculation of distance between 2 phonetic strings.

It is interesting to mention that such a method could be useful to evaluate, in a precise lexicon, the degree of difficulty of decoding: a great distance between words could indicate an easy task of decoding.

We also mention that phonological rules are necessary to forecast all the variations of a word's pronunciation. For example, the word "petite" (French word for "small"), whose phonological transcription is /pətitə/, can be pronounced as [ptit], [pətit], [ptitə] or [pətitə] and should have 4 phonetic entries in the dictionary.

RESULTS

In a first step, each module has been tested independently using 10 French speakers (5 male, 5 female) recorded in the corpus SYL of the French database BD-SONS. Stimuli were CVCV non-words as "titi", "rara", "sussu"... that represents the combination of 2 * 10 vowels * 16 consonants * 10 speakers = 3200 tests.

Results of the global recognition

For each tested stimulus, the module can provide different candidates. We distinguish (Table 1): the case where the good phoneme is in the list of candidates (column *'Correct'*) and the case where it is not (column *'Error'*). The column *'Num cand'* furnishes the average number of candidates proposed by the module, which can be compared with the number of potential candidates (column *'Num max'*). For the vowels, the identification process classifies the candidates: the column *'1st rank'* (Table 1b) is the case where the good vowel is in the first position, the column *'other rank'* is the case where the good vowel is in the second, the third, ... position.

Table1 : Results of the global recognition

(a)	Correct	Error	Num cand	Num max
Consonants	86 %	14 %	2.89	17

	Correct				
(b)	1st rank	other rank	Error	Num cand	Num max
Vowels	65 %	25 %	10 %	2.18	10
	90 %				

Results of the analytic recognition

Here, we are only presenting the tests relative to the identification of vowels.

In the module of analytic recognition by rules, the system proposes 2.35 candidates on average, among the ten vowels [a,i,u,o,e,y,ø,ɑ̃,ɛ̃,ɔ̃]. In 92.4% of the cases, the good vowel is among the candidates.

The module of analytic recognition by networks has been tested using a corpus of 42 words pronounced by 5 speakers. On about 450 vowels, the result was correct in 92% of the cases.

CONCLUSION

We have presented the different parts of a system based on parallel distributed processes for speaker independent acoustic-phonetic decoding. Each module seems efficient in the recognition task.

Our aim now is to integrate all these knowledge-sources to collaborate in a single system. We are testing it with a large corpus of 500 French words pronounced by seven speakers. The results will be published in a next paper.

REFERENCES

[1] Lindsay P., Norman D. (1977), *Human information processing - An introduction to psychology,* Academic Press, New York, USA.

[2] Edelman G.M (1992), *Bright air, Brilliant Fires : on the matter of mind,* Basic books, New York, USA.

[3] Meyers C.S., Rabiner L.R. (1981), "A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition", *IEEE ASSP*, vol.29, pp. 284-297.

[4] Rossi M. (1990), "Automatic speech segmentation:why and what segments ?", *Revue Traitement du Signal*, GRETSI, vol.7, n°4, pp. 315-326

[5] Ghio A., Rossi M. (1994), "Reconnaissance globale et analytique dans SYMULDEPHO, un système multi-locuteurs de décodage acoustico-phonétique", *Proceedings of workshop on 'Automatic Speech Recognition'* Nancy,France, GDR-PRC Man-Machine Communication

[6] Yong G., Mason J.S. (1987), "A comparison beetween vocal tract and auditory feature analysis in ASR.", *Proceedings of Eurospeech*, Edinburgh, pp.132-135

[7] Hermansky H. (1990), "Perceptual linear predictive (PLP) analysis of speech.", *J.Acous.Soc.Am.*, vol.87, *pp.* 1738-1752.

[8] Wagner R.A, Fisher M.J. (1974), "The string-to-string correction problem", *Journal of the ACM*, 21,1.

[9] Véronis J.(1994), "Distance entre chaînes: extension aux erreurs phonographiques", *Travaux de l'Institut de Phonétique d'Aix*, vol.15,pp.219-233.

CONTRASTIVE ACCENTS – HOW TO GET THEM AND WHAT DO THEY LOOK LIKE

J. Haas[1], *A. Kießling*[1], *E. Nöth*[1], *H. Niemann*[1], *A. Batliner*[2],
[1] *Lehrstuhl f. Mustererkennung (Inf. 5), Universität Erlangen-Nürnberg, Erlangen, FRG*
[2] *Institut f. Deutsche Philologie, L.M.-Universität München, München, FRG*

ABSTRACT

Automatic dialog systems tested with naive users are often confronted with special speaking styles, as e.g. words produced with emphatic or contrastive accent. Such utterances usually cause problems for word recognizers, because they were not included in the training data. It is thus important for the improvement of future systems to be able to collect utterances containing contrastive accents produced as natural as possible. We describe in this paper an automatic simulation system for provoking and collecting contrastive accents. With this system, 15 recording sessions were conducted; in total 205 word tokens produced either with default or with contrastive accent were collected. We discuss the results of an automatic classification as well as the relevance of extracted prosodic features for the marking of contrastive accents.

INTRODUCTION

While testing our automatic speech understanding and dialog system EVAR with naive users (via public telephone) the following situation was often observed: Because parts of the user utterance are not recognized correctly, the system delivers the wrong information. Usually, the user repeats the misrecognized words in a special, often excessive manner, using emphatic or contrastive accent. These utterances cause all the more recognition problems (not only for EVAR, but for all existing word recognition systems), because they were not included in the training data, and the dialog fails. Thus, there is a strong need for the collection of utterances produced with emphatic or contrastive accents and to take them into consideration during the training phase.

For the collection of words or phrases with contrastive accent it is essential that the data are produced as natural as possible. Asking speakers to read contrastive accents is a traditional [1] but suboptimal way. On the other hand, spontaneous speech corpora from human–human–dialogs contain very few contrastive accents. For example, in 20 investigated dialogs (approx. 60 min speech) of the VERBMOBIL–Corpus [4] no single contrastive accent could be observed. Another possibility for the collection of contrastive accents is to use the human–machine–dialogs conducted with the EVAR system. However, compared to all user utterances the occurrence of contrastive accents is not that high, and therefore very much effort had to be put on their identification.

In this paper we describe an automatic system with which a large amount of naturally produced contrastive accents can be provoked and collected. The system conducts dialogs with naive users by simulating an automatic speech understanding system in the domain of "train time table inquiries". It is designed to collect prosodic minimal pairs of words containing either the default word accent or a contrastive accent. In the second case, the position of the contrastive accent (either on the lexical word accent syllable or on a different one) can be induced by the system. It is thus possible to overcome the paradox to provoke spontaneously produced prosodic minimal pairs in an experimental environment.

THE SIMULATION SYSTEM

The simulation system is a Wizard–of–Oz–System where the role of the human wizard is played by the machine. Because it is no human wizard who can react on any possible user utterance in a flexible manner but a simple computer program, the structure of each dialog conducted with the simulation system is heavily restricted. In any state of the dialog, the system has to react in such a way that there is no other possibility for the user than to behave in an expected, predefined manner. On the other hand, it is essential to prevent the user from realizing that he/she is *not* communicating with a 'normal' automatic system. One way of doing this is to produce a well-balanced proportion of 'artificial' recognition errors in the system's output. The speaking style of the users should not be influenced, and therefore, the output of the system is always presented in textual form on the screen; no synthesized speech is used. To prevent the users of becoming bored too soon and to get as natural utterances as possible it is important to provide them with a good amount of different alternating system reactions as well as to grant them from time to time a sense of achievement by passing the correct train time table information right after the first query.

For all these reasons, much care had to be put on the design of the system. Additionally, to be aware of any other unforeseen problem, each recording session can be accompanied by a supervising person that knows about the structure of the simulation system and can guide the user in the right direction.

The first and very important step to guide the user into the predefined dialog is to start each dialog with a train time table inquiry given on the screen to be read by the user, e.g:

U: *"I want to go to Hamburg."*

From these first queries the tokens for the default accents (→ target_D1) were collected. After this query different system reactions are possible (cf. Figure 1), each of them provoking a specific user reaction. In the first situation (1a) a correct recognition of the user's query is simulated and the requested information, i.e. the correct train connection, is given. This provokes no specific user reaction but grants him/her a feeling of success. In (1b) a correct recognition is simulated, asking the user for confirmation. The usually following single word utterances (e.g. *"yes"*) or any other type of confirmation, can be collected as a by-product and re-used for training.

1. System simulates correct recognition of the user utterance
   (a) System does not ask back (i.e. passing desired information)
   S: *"You can take the train at 10.47 ..."*
   (b) System asks back
   S: *"You want to go to Hamburg?"*
2. System simulates recognition error
   (a) System provokes contrastive accent on the word accent syllable
   S: *"Do you want to go to Hamburg or to Homburg?"*
   (b) System provokes contrastive accent on the second syllable
   S: *"You want to go to Hamberg?"*
   (c) System provokes a distinct (emphatic) pronunciation
   S: *"Where do you want to go?"*

Figure 1: *Possible system reactions following the first user query.*

The best way to provoke the user to put stress on a specific syllable is to simulate recognition errors. In (2a) the user is induced to produce a contrastive accent on the lexical word accent position (usually he/she's going to utter: *"To* **Ham***burg"*). These utterances are used to collect the first type of contrastive accent (→ target_C1). With system reaction (2b) a contrastive accent on a specific syllable different from the word accent syllable can be provoked (induced user utterance: *"No, to Ham*burg"). In this case the stress is put on the second syllable of the word (→ target_C2). With system reaction (2c) the user is induced to use a very distinct (emphatic) pronunciation where sometimes both syllables (→ target_C12) are overemphasized (esp. if this situation is used several times subsequently). Note that this mode of provoking accents was

not used for the words examined in the following.

EXPERIMENTS AND RESULTS

Using the simulation system 15 recording sessions with 15 different users (180 dialogs in total) were conducted for collecting different types of accentuations, where all the intended minimal pairs comprised city names (like *"Hamburg"*, *"Freiburg"*) or time expressions (like *"at nine o'clock"*). Most of the users were students from the computer science department with no special knowledge of speech recognition or the EVAR system. They were told that their task is to test the automatic speech understanding system, and for the sake of convenience for the transcriber the first user utterance has to be read from the screen. At the end of each recording session, the users were asked about their experience with the system. None of them had any doubt that he/she was working with an automatic dialog system; most of them were very surprised about the systems capabilities and the computational speed.

In total 205 word tokens were collected, recorded and digitized using a *Desklab 14* from *Gradient*. Most of the tokens (62) were obtained for the city name *Hamburg*; in the following discussion, we confine ourselves to these items. The tokens were cut out of the signal, the syllable boundaries were adjusted by automatic time-alignment using an HMM-based word recognizer and corrected manually.

In an informal perceptual evaluation it was checked that the induced accentuation types were produced in the expected manner. Only 6 % of the induced contrastive accents were perceived as default accent; none of the default accents was perceived as a contrastive accent.

For the investigation of the prosodic properties of the different induced accentuation types, F0-contour and rms-energy (frame length: 10 ms) were computed automatically using the algorithm described in [3]. The F0-values were transformed into semi-tones. For F0 and energy the mean over the whole word was subtracted from each value. The following prosodic features were computed for each syllable: minimum, maximum, range, mean, onset and offset of the F0-contour; duration of the syllable nucleus; mean of the energy-contour.

Table 1: *Confusion matrix of induced and automatically classified accentuation types in percent.*

	# Tk.	D1	C1	C2
target_D1	28	78.6	10.7	10.7
target_C1	19	5.3	84.2	10.5
target_C2	15	13.3	26.7	60.0

In Table 1, the result of an automatic classification is shown (linear discriminant analysis, learn = test, all features used in a forced entry design). At first sight, the low recognition rate for target_C2 might surprise: 60% correct, and 26.7% confusion not with the default case target_D1 but with target_C1 where an 'opposite' accent pattern is expected. Of course, misproductions cannot be ruled out altogether and might – esp. if the number of tokens is as low as in our case – heavily influence the classification results. A systematic explanation along the lines of [2] can, however, be offered. There, a double focus on two different words was induced by the context but often it was classified and perceived *not* with focal accents on these two words but with one single accent on the word in the default ("out of the blue") accent position. But that means that speakers who do not "behave properly" – i.e. as the linguist likes them to do – do nevertheless deviate in a systematic manner. The same might be the case with contrastive accents: The strategy of naive speakers when confronted with a "contrastive misunderstanding" (*Hamberg*") instead of *Hamburg*) might sometimes be simply to repeat the word in question more pronounced in an overall manner but *not* – or not only – with a contrastive accent on the misunderstood syllable. As far as this behavior is representative for real life applications, it must be accounted for in the system.

In Table 2 the average of the feature values for both syllables is shown for the three induced classes. The duration of the syllable nucleus is most significant for distinguishing default from contrastive accent; the tokens with contrastive accent are clearly longer than the default accents. The ratios between first and second syllable for default accent (1.29), contrastive accent on the first syllable (1.23) and contrast on the second syllable (1.14) moves towards a comparatively longer second syllable with the weakest differences in total syllable nucleus duration for target_C2. Still, the mean value of the absolute duration of the first syllable is for target_C2 slightly longer than for target_C1 and this fact corroborates our hypothesis that contrastive accentuation is not strictly refined to the syllable in question. The difference between the F0 features is not that distinct. The F0-range on the second syllable is clearly smaller for the default accent; the F0-mean, however, rises from the first to the second syllable. The energy proportions between first and second syllable show high differences for all three accentuation types. For the contrastive accents, these differences are as expected: higher energy on the accentuated syllable. For the default case, it is the other way round. Possible reasons might be that target_D1 was embedded in a complete sentence whereas the contrastive accents were usually just one word utterances and that no phoneme intrinsic normalization was performed for the energy.

The same features were extracted also for the automatically determined (not manually corrected) syllable positions. Same tendencies in the feature behavior could be observed, the differences were, however, less distinct.

Table 2: *Average feature values for the three induced classes.*

Feature	target_D1		target_C1		target_C2	
# Token	28		19		15	
	Syl. 1	Syl. 2	Syl. 1	Syl. 2	Syl. 1	Syl. 2
Nucleus duration	153.6	119.2	192.4	156.0	201.0	176.5
F0–Mean	-0.15	0.37	0.40	-0.43	0.55	-0.55
F0–Maximum	1.61	1.75	2.16	1.32	2.60	1.60
F0–Minimum	-2.04	-1.25	-1.42	-2.05	-1.60	-2.93
F0–Range	3.64	3.00	3.58	3.37	4.20	4.53
F0–Onset	-0.39	0.11	-0.63	0.79	-0.53	0.13
F0–Offset	0.00	-0.07	0.89	-1.53	0.33	-1.87
Energy–Mean	-7.34	12.49	6.16	-4.48	-2.41	2.74

CONCLUDING REMARKS

It has been shown that with the system described here, an automatic collection of contrastive accents produced in a natural way can easily be performed. Not only contrastive accents can by provoked with the system but, with some slight modifications of the system design, also other spontaneous speech phenomena like hesitations. Furthermore, preliminary experiments have already been conducted for the collection of spontaneous speech phenomena with the so called *"shocking effect"*, where an absolutely unexpected system answer like *"Why do you want to go there?"* is provoking very surprised user reactions.

ACKNOWLEDGEMENT

This work was funded by the German Federal Ministry of Education, Science, Research and Technology (BMBF) in the framework of the Verbmobil Project under Grants 01 IV 102 F/4 and 01 IV 102 H/0. The responsibility for the contents lies with the authors.

REFERENCES

[1] R. Bannert. Fokus, Kontrast und Phrasenintonation im Deutschen. *Zeitschrift für Dialektologie und Linguistik*, Vol. 52, pp. 289–305, 1985.

[2] A. Batliner, W. Oppenrieder, E. Nöth, and G. Stallwitz. The Intonational Marking of Focal Structure: Wishful Thinking or Hard Fact? In *Proc. XIIth ICPhS*, Vol. 3, pp. 278–281, 1991.

[3] A. Kießling, R. Kompe, H. Niemann, E. Nöth, and A. Batliner. DP-Based Determination of *F0* contours from speech signals. In *Proc. ICASSP'92*, Vol. 2, pp. 17–20, 1992.

[4] W. Wahlster. Verbmobil — Translation of Face-To-Face Dialogs. In *Proc. EUROSPEECH'93*, "Opening and Plenary Sessions", pp. 29–38, 1993.

PHONETIC CHARACTERISATION AND LEXICAL ACCESS IN NON-SEGMENTAL SPEECH RECOGNITION

Mark Huckvale
University College London, London, U.K.

ABSTRACT

An isolated-word speech recognition system, built without the use of linear segments for acoustic modelling or lexical access, is justified, described and demonstrated. The system comprises phonetic feature analysis operating on four independent tiers, parallel phonotactic parsing, and lexical access based on a neural-network inspired lexicon structure. Performance is however still inferior to a baseline segmental system.

INTRODUCTION

This paper describes an attempt to bring together into a single operational system a selection of alternatives to the linear segmental approach to phonetic modelling and lexical access found in contemporary automatic speech recognition systems.

The most important departure from current architectures is the explicit separation of phonetics and phonology in the system. In the new system the role of the first is to characterise speech-specific elements of the sound signal, while the role of the second is to establish the functions of these elements in linguistic encoding. In contrast, current systems based on phones-in-context use linear phonological units to organise their acoustic models as well as for lexical access. Such systems have particular weaknesses, including (i) poor modelling of variation of acoustic realisation of phonological units in context, (ii) failure to model post-lexical phonetic variety because of the need for complex and arbitrary context-sensitive realisation rules, (iii) failure to exploit contextual variation as discriminative information, (iv) failure to use temporally extended information relevant to phonological identity, (v) failure to exploit prosodic structure in the signal. These weaknesses lead to systems which lack discriminative power, are unable to exploit known pronunciation variety in context or in accent, fail to extract the most from impoverished signals, and ignore the information and constraints available in the rhythm, stress and intonation of the speech.

On the other hand, linear phonological-unit based acoustic models provide a simple and computationally effective basis for recognition. There is a synergy between a linear phonological account and syntactic pattern recognition algorithms such as Hidden Markov Modelling (particularly the Viterbi decoding scheme). It has been said that in speech recognition good knowledge is of no use without good algorithms for applying it. Hidden Markov Modelling has been successful because it forms a *coherent* view of the acoustic to phonological mapping, rather than an accurate one.

Thus the challenge is to find effective procedures for the exploitation of more sophisticated models of speech.

DESIGN

In this section we justify the non-segmental recognition system described in the following section. More details may be found in [1].

Phonetic component

The role of the phonetic component in a non-segmental system is to model the range of variety of acoustic realisation of elemental phonetic characteristics. For each given characteristic at each time frame, the phonetic component supplies the probability that the element has been realised (by a given speaker in a given acoustic environment). By relaxing the requirement that these characteristics need to be themselves phonological we can make this component more sensitive to sub-phonemic changes, to syllabic and prosodic structure. Although we can no longer exploit phonological sequence constraints we can still exploit phonetic constructional constraints that arise due to the fact that the signal was spoken. In the simplest model, the phonetic component operates on a number of *tiers* where the phonetic properties inside a tier are mutually exclusive, while properties across tiers are mutually independent. As we shall see this allows the use of a syntactic pattern recognition scheme to operate within a tier.

Lexical Access

From the phonetic characterisation of the signal it is necessary to explain the phonetic evidence as realisations of a sequence of words, subject to a number of constraints: (i) words occur strictly sequentially, (i.e. only one word is active at any one time), (ii) citation form phonetic structures of words are subject to a limited range of contextual modifications, (iii) word selection is guided by the task (vocabulary, syntax, etc.).

Since at this stage we do not have a phonological representation, all we can do is activate word hypotheses on the basis of the likelihood that they might have given rise to the phonetic evidence. Following the TRACE model of lexical access [2] we can see that each phonetic characteristic can feed 'activation' into the lexicon, (but in this case without an interposing phonemic layer). Given a tiered phonetic analysis, any single tier activates a number of possible word hypotheses. The initial activations of words need not be zero, since there may be prior evidence (from the task) for the likelihood of words.

Phonological categorisation

From the word activations (over time), it is necessary to determine the most likely word sequence. Unfortunately, what we have at the moment is essentially a whole-word template recognition system, and it is easy to show that such systems cannot be extended to large vocabularies without the exploitation of *phonological* knowledge. Each word has been activated on the basis of phonetic similarity with the input, but it is likely that some components of the word match better than other components. Thus the vowel of [pi] may match the input quite well, but the consonant may match badly. If each word has independent pronunciation models of phonetic realisation, it is possible that the vowel of [ti] might not match as well as the vowel of [pi]. Thus an input '"T" may be recognised as "P" because the vowel matches overcome the consonantal matches. The solution to this is to indicate that the vowel in [pi], i.e. /i/, is the same as the vowel /i/ in [ti]. With this constraint, the difference in the vowel scores is irrelevant and the consonantal match controls the outcome. This is *phonological* knowledge that must be specified in addition to the phonetic realisation of words.

One way of imposing these phonological constraints is to establish a set of phonological units above the words, which share activations between words which have similar phonological prescriptions. Thus an /i/ unit short-circuits activation between [pi] and [ti] to counteract exactly any difference due to independent models of the vowel.

ARCHITECTURE

The specific implementation of the non-segmental recognition architecture for an isolated word recognition task may be separated into: (i) multiple Phonetic feature components that deliver phonetic feature analyses of 30ms of speech signal, (ii) Phonotactic decoding components that deliver element sequence likelihoods for each tier, (iii) a

Lexical access component that takes the element sequence scores and delivers a word hypothesis using lexical and phonological information. More details may be found in [3].

Phonetic feature component

The phonetic feature component operates on four independent tiers, corresponding to multiple broad-class analyses of the signal.

In the *Excitation* tier, phonetic elements represent Silence (SIL), Voicing (VOI), Frication (FRC) and Mixed excitation (MIX). In the *Degree* tier, elements represent Oral closure (STP), Nasal (NAS), Fricative (FRC), Approximant (APP), Close vowel (CLS), Mid Vowel (MID) and Open vowel (OPN). In the *Position* tier, elements represent Labial (LAB), Dental (DEN), Alveolar excluding /s/ (ALV), /s/ frication (FRS), Front/Palatal (FRN), Central (CEN), Back (BAK), Velar (VEL) and Silence (SIL). In the *Strength* tier, the elements represent Burst (BUR), Aspiration (ASP), Other frication (FRC), Vocalic (VOW), Voiced plosive (VGP) and Silence (SIL).

These tiers together are sufficient to differentiate English words apart from short and long vowels at a single place (e.g. bit vs. beat) and dental and labiodental fricatives (e.g. thin vs. fin). Performance on elements for these contrasts is currently unsatisfactory.

For each tier, a Multi-Layer Perceptron (MLP) classifier was trained between a spectral representation of the signal and the target element classes. Each tier had its own MLP with 3x10ms frames representing 19 filterbank energies + overall energy (i.e. 60 parameters) as input and 1 output per element class. Each MLP had a single hidden layer of a size equal to three times the output layer size. The training data was 666 different monosyllabic words spoken by one speaker. There were approximately 83,000 training vectors.

Each training word was annotated and the element labels generated by rule using a mapping that took into account boundaries and the nature of adjoining segments. Training was performed using an adaptive back-propagation method firstly on the automatically generated element labels, and then, after realignment with the partially trained network, against realigned element labels.

Phonotactic decoding component

To generate an element sequence, a Viterbi decoding was performed on the MLP outputs for a tier over the whole duration of a word. See Figure 1. This process delivered a score for each phonotactically possible sequence in the test vocabulary for each tier. Over the 4 tiers there were 450 possible element sequences, but only the best scoring 50% in each tier were used for lexical access

Lexical access component

To identify the lexical item a network lexicon was used based on [1]. Here the phonetic input was provided by the element sequence scores; these then fed activations to the word units according to 'dictionary' pronunciations of the words. Thus words were only connected to element sequences expected in the citation pronunciation. To smooth activations across words, a level of phonological units were constructed above the word units, which channelled activation between words sharing similar phonological descriptions - in this experiment shared syllabic components. Thus word activations arose primarily from the phonetic input, but subsequently there was interaction and competition between words mediated by a set of phonological units. The most strongly activated word unit was chosen to be the recognised word.

RESULTS

For testing the architecture, 359 monosyllabic words, different to the training words, but spoken by the same speaker were used. The raw recognition performance of the Phonetic feature analysis component was:

Tier	Frames correct
Excitation	91.6 %
Degree	82.8 %
Position	74.7 %
Strength	87.9 %

The raw recognition scores for the element sequences was:

Tier	Top 1	Top 5
Excitation	76.6 %	98.3 %
Degree	46.0 %	80.2 %
Position	23.4 %	52.9 %
Strength	44.0 %	88.6 %

For the feature-to-word activations alone, without the use of the phonological units for smoothing, the word recognition performance was 51%. Small amounts of phonological unit activation fed back to the word units improved recognition performance only slightly, to 53%. Performance is so weak primarily due to the poor performance of the Position tier.

Baseline recognition performance using a monophone HMM trained on the same material (and having approximately the same number of free parameters as the set of MLPs) was over 90%.

FURTHER INFORMATION

The author welcomes comments on `M.Huckvale@ucl.ac.uk`

REFERENCES

[1] Huckvale, M.A. (1990), "Exploiting Speech Knowledge in Neural Nets for Recognition", *Speech Communication*, p1.

[2] McClelland, J.E. & Elman, J.L. (1986), "Interactive Processes in Speech Perception: The TRACE Model", in *Parallel Distributed Processing*, ed Rumelhart & McClelland, MIT Press.

[3] Huckvale, M.A. (1994), "Word Recognition from Tiered Phonological Models", *Proc. IOA Conf. Speech and Hearing*, Windermere.

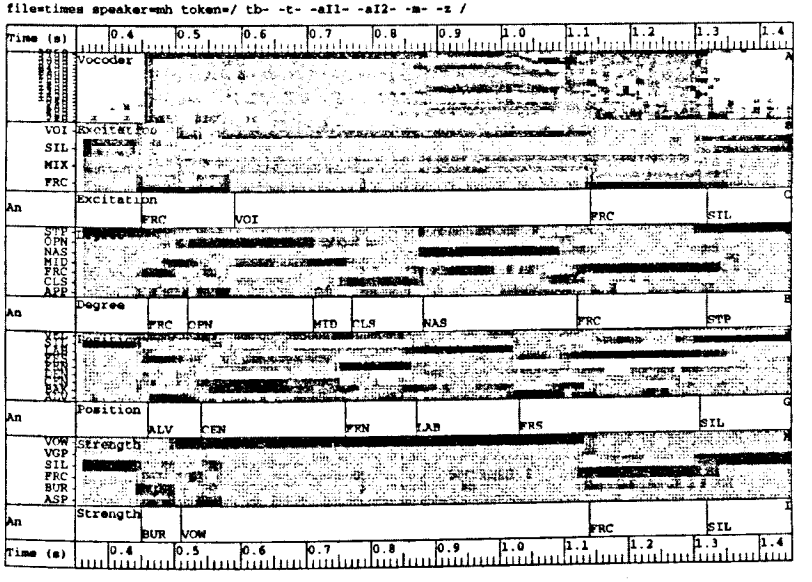


Figure 1. Tiered analysis of the test word 'times'.

IMPROVING SPEECH RECOGNITION WITH MULTIMODAL ARTICULATORY ACOUSTIC HMMs

B. Jacob - C. Sénac - R. André-Obrecht - F. Pellegrino
IRIT- Université Paul Sabatier - CNRS URA 1399
118,route de Narbonne
31062 Toulouse Cedex FRANCE

ABSTRACT

This paper describes a new scheme for robust speech recognition systems where visual information and acoustic features are merged. A segmental processing and two decoding strategies based on Hidden Markov Models (HMM), are studied and evaluated on connected word recognition applications.

INTRODUCTION

The proposed recognition system is one of the components of the AMIBE project (Applications Multimodales pour Interfaces et Bornes Evoluées). The purpose of this work, supported by the PRC's Informatique (Coordinated Research Programs of the CNRS) is to study the natural visual and auditive bi-modality of oral communication and to propose more robust speaker verification and speech recognition systems.

It's well known that, listening in adverse acoustic environments (noise, multiple speakers...) relies heavily on the visual input to disambiguate among acoustically confusable speech elements [1]. To take this phenomena into account, we develop an Automatic Speech Recognition system which processes the synchronization of the 'labial reading' and an acoustic pattern recognition system using HMM.

The lip-reading consists in a pre-processing of the visual information, thus producing a set of articulatory features as described in [2]. The acoustic pre-processing is based on a segmentation algorithm followed by a cepstral analysis. But as articulatory target positions and acoustic steady segments are not always synchronized, we propose two different strategies for merging these two kinds of data:

- a concatenation of the cepstral and labial vectors which provides a global observation vector for a classical HMM;
- a master/slave type relationship between two HMMs [3] which leads to correlate the two informations.

RECOGNIZER OVERVIEW

An automatic speech recognition system involves basically two components : the prepocessing to reduce the information and the linguistic decoder.

Extraction of the signal parameters

Our recognition system processes two kinds of signal :
- an acoustic signal sampled at 16 kHz,
- three articulatory signals composed of the lip breadth (A), the lip height (B) and the lip area (S), sampled at 50Hz [4].

The acoustic signal is preprocessed by an automatic segmentation [5] and a spectral analysis is performed on each segment. Therefore 8 Mel frequency cepstral coefficients (MFCC) are extracted. We add the energy (E) and their first derivatives (8 δMFCC, δE).

The acoustic segment boundaries are projected on the articulatory signals; for each segment, we calculate the mean of each labial parameter and their first derivatives.

The global feature vector consists of 18 acoustic coefficients, 6 articulatory ones, and the duration of the segment (T). Figure 1 gives an example of an acoustic signal preprocessed by the automatic segmentation

Statistic models of the linguistic decoder

Two different approaches are proposed :

- a global standard H.M.M., M_{glob}, is hierarchically built ; each word model is obtained by concatenation of elementary acoustic models. The elementary unit is the 'pseudo-diphone' ; it corresponds to the steady part of a phone or the transient parts between adjacent sounds and the acoustic model is a basic left to right continuous density H.M.M. ;
- in the master/slave approach, two parallel H.M.Ms are built. The first one, named articulatory H.M.M. M_{art}, is an ergodic model of three states and three pdfs, which takes the articulatory features into account. The second one, named acoustic H.M.M. M_{acous}, has the same topology as M_{glob} and processes the acoustic observations only. The M_{art} H.M.M. controls the M_{acous} HMM, in the sense that the M_{acous} H.M.M. transition and observation probabilities depend on the current state in M_{art} [3].

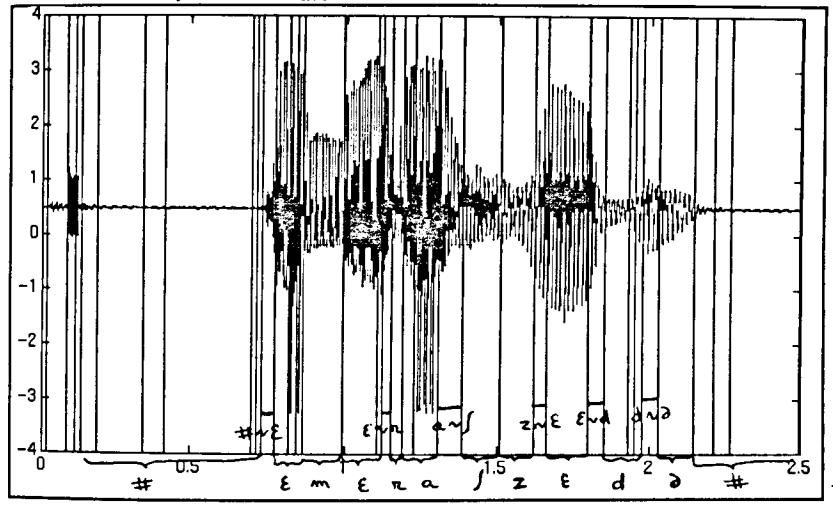


Figure 1: Example of a preprocessed acoustic signal by automatic segmentation. The four spelled letters "M R H Z" are pronounced. For each segment, is indicated the pseudo diphone label found by the Viterbi algorithm, with the M_{glob} model :
 # represents 'silence';
 # ~ɛ , ɛ ~ m, ɛ ~ r, a ~ʃ ... represent transient units,
a, d ... represent steady units.

EXPERIMENTATIONS

We have experimented these two approaches on two mono speaker applications : connected digit recognition and connected spelled letter recognition. The connected digit corpus is composed of sequences of four digits : 84 sentences for the learning set and 35 sentences for the test set. The connected spelled letter database is composed of sequences of four spelled letters : 158 sequences for the learning set and 48 for the test set.

Table 1 gives the error numbers on the digit test set, in terms of sentences and words. It is well known that very good results are obtained with a classical HMM such as M_{glob} and with 8 standard MFCCs. We observe no performance loss when using the segmental processing (1 word substitution) and a very small one when introducing the labial information (3 word substitutions). The comparison between the classical HMM M_{glob} and the master/slave $M_{acous}+M_{art}$, shows a better result for the global approach (3 errors vs 5 errors), but this remark must be qualify : first the confidence interval doesn't permit a precise conclusion, and then the complexity of the master/slave HMM is such that the number of parameters is too important to hope a good learning with such a little learning set

Table1: Recognition error number, in terms of sentences and words, in accordance with the parameters and the models.

Model	Coefficients	Sentences /35	Words /125
M_{glob}	8MFCC+E+T	1	1
	8MFCC+E+T +A+B+S	3	3
master/slave $M_{acous} + M_{art}$	8MFCC+E+T +A+B+S	5	5

For the spelled letter application, an initial M_{glob} model is learned with 8 MFCC, the energy and the segment duration ; we add successively the 3 labial parameters and their derivatives. The same experiment is repeated with an initial global HMM learned with 8 MFCC, the four first derivatives, the energy and its derivative, and the segment duration. The results are reported on Figure 2. We observe that the best recognition rate is obtained when using the lip height and breadth. The introduction of the lip area doesn't bring any pertinent information, it is strongly correlated with the parameters A and B. The derivatives appear as noisy information, the desynchronization between the labial information and the acoustic one is certainly one of the cause. On Figure 1, we can see the alignment of the sentence "M R H Z" obtained by the Viterbi algorithm through the best M_{glob} model (8MFCC, E, T, A, B), in terms of pseudo-diphone units ; segments and pseudo diphone units are perfectly aligned.

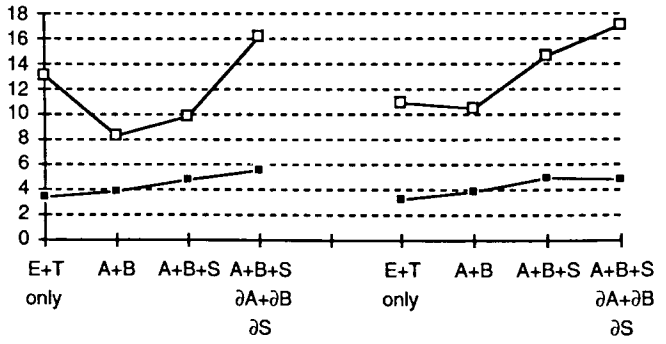


Figure 2: Recognition with the M_{glob} HMM : Word error rate for the learning set (■) and for the test set (□) in accordance with the articulatory vector. On the left part of the figure, the acoustic vector consists of 8 MFCC, E, T as on the right part, the first four derivatives are added.

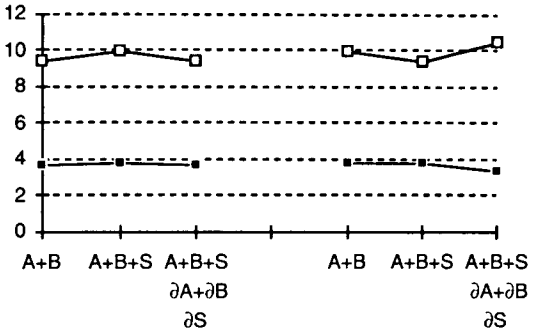


Figure 3: Master/slave M_{acous} +M_{art} HMM results : word error rate for the learning set (■) and for the test set (□) in accordance of with the articulatory vector. On the left part of the figure, the acoustic vector consists of 8 MFCC, E, T as on the right part, the first four derivatives are added.

The same experimental protocol is performed to test the master/slave approach ; of course, we add to the initial parameters the labial ones, A and B. The results are reported on the Figure 3. We notice that the results are quite so good as using the global approach and they are more steady : the introduction of supplementary parameters (labial or acoustic derivatives) doesn't disturb the recognition rate , in view of the confidence interval. This remark is very promising, we may hope that this structure is the best one to introduce the labial information and that it will be more robust when the acoustic parameters will be noisy. Future experiments must confirm this conclusion.

CONCLUSION

We have described two statistical approaches based on HMMs to merge articulatory and acoustic information and to improve an automatic speech recognition. Experimental results show the difficulty to process the desynchronization between the lip moves and the acoustic signal. It seems that the more robust approach is the master/slave one, future studies must confirm this hypothesis.

REFERENCES

[1] P.Duchnowski, Meier U., Waibel A. : "See me, hear me: integrating automatic speech recognition and lip-reading". S11-6.1 ICSLP 94, YOKOHAMA.

[2] C. Benoit , Abry C., Boe L.J. : "The effect of context on labiality in french". Eurospeech 91, GENOVA.

[3] F. Brugnara, de Mori R., Guiliani D., Omologo M. : "A family of parallel HMM". p 1103, Proc.IEEE Int. Conf. ASSP 92, SAN FRANCISCO.

[4] M.T. Lallouache: "Un poste "visage parole" couleur. Acquisition et traitement automatique des contours de lèvres". Thèse de Doctorat de l'Institut National Polytechnique de Grenoble, 1991.

[5] R. André-Obrecht: "A new statistical approach for the automatic segmentation of continuous speech signals". IEEE Trans. on Acoustics, Speech, Signal Processing, vol 36 n°1, January 1988.

INTEGRATING VISUAL AND ACOUSTIC INFORMATIONS IN A SPEECH RECOGNITION SYSTEM BASED ON HMM

P. Jourlin, M. El-Bèze and H. Méloni
Laboratoire d'informatique, Avignon, France

ABSTRACT

In an unprotected sonorous environment, an ever so slightly high level of noise can be a hindrance to the correct perception of the acoustic message. Works on multimodal perception show how visual information leads to a compensation of information losses caused by acoustic noises. This explains why an HMM-based system which is able to take information from both visual and acoustic sources can be interesting for speech recognition in noise.

INTRODUCTION

Two main goals appear in speech recognition research : reduce the number of constraints upon working conditions and improve recognition rates of the systems.

The integration of other information sources is actually more and more taken into consideration. These can be a knowledge on different levels : articulatory, auditory, syntactic, morphological, semantic, and so on.

But we can also contemplate taking into account visual information which accompanies, even determines the sound emission. The lips take part in this speech production process.

So, we can feel free to think that labial movements can help speech recognition improvement.

SITUATION OF THE PROBLEM

The main object of this paper is to create an automatic speech recognition system which can draw benefit from lips-originated information.

It is within the AMIBE PROJECT context (Applications Multimodales pour Interfaces et Bornes Evoluées) for which one of the applications could be the conception of advanced bank interfaces.

In this framework, the acoustic signal may be noisy without the video signal being so. This is why labial movements can help to compensate the loss of information due to noise. Some studies have brought to the fore this contribution amongst various persons as well as some problems :

- The labial anticipation and retention which creates a disynchronization between visual and acoustic information, [1],[4].
- The existence of labial doubles which limits labial recognition to some groups of phonemes,[1] [6], [7].
- Coarticulation effects, the incidence of which upon speech signal have been widely studied, [3].

The systems described below are realized from HTK (HMM toolkit) of the C.U.E.D. (Cambridge University Engineering Department Speech Group), some experiments having need a modification of sources.

HIDDEN MARKOV MODELS

Introduction

We chose to resort to these probability models because their use does not require, at first, a thorough knowledge in the application field. The learning stage enable us to detect automatically significant information from the data to which they are submitted.

Each model is represented by a Markov *source* (see Figure 1), which is a probabilistic automaton of finite states, which means that each transition has a probability to be used and a probability to emit symbols.

[1] Also called visemes

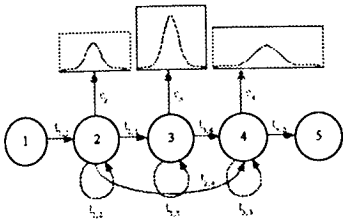


Figure 1 : A Markov source

In the field of speech, these symbols are *N*-dimensional acoustic vectors (ex : LPC of *N* coefficients) but they may also be labial vectors. In the case of a continuous model, the probability emission of these vectors is ruled by a Gaussian mixture.

When some vector components are supposed statistically independents, we can split them into several *streams* (we have one probability distribution for each state and each stream), the calculation of the probability of emission being the product of the probabilities for each stream. The emission and transition probabilities of each model will be estimated from learning data.

Learning stage

We have a set of parametrized and labelled signals at our disposal. The learning stage is composed of 3 stages.

Initialization

Using the *K-means* classification method, we associate to each state of each model, the vectors stemming from the various examples of the concerned word. It needs providing a set of labelled data as an entry of the process.

The start of the learning phase induces a manual intervention. Thanks to the maximum likelihood estimator, we even initialized for each state the means and variances of each component of these vectors, which will enable us to get the associated probability distribution.

This method realizes a rough state-level segmentation as it does not take into account the probabilities of transition and leaves asides the sequential order of the vectors. These various limitations call for the use of a second stage.

Probabilities estimation with temporal constraints

We compute the transition probabilities values by using the Baum-Welch algorithm (also called Forward-Backward). It appears that this stage, based on the initial boundaries of words pains into managing into the questions of contexts related to continuous speech.

This is why it is necessary to proceed a third stage.

Probabilities estimation without temporal constraints

For each sentence of learning data, we create a concatenation of the models corresponding to words pronounced.

Then, we use the Baum-Welch algorithm on the sentence and the associated model, which enables us to take into account the effects related to the context for each element of the concatenation.

This last step enables us to discuss the boundaries of labelling.

Decoding

To recognize the underlying word associated with a given sequence of vectors, we compute for each model the probability that it may have emitted it : the word recognized corresponds then to the model maximizing this last value.

In other words, if O is the sequence of vectors observed, we must compute :

$$\arg\max_i P(w_i \mid O)$$

This value is not computable directly but using Bayes'rule gives :

$$P(w_i \mid O) = \frac{P(O \mid w_i) P(w_i)}{P(O)}$$

w_i being the i-th word of the vocabulary and O the studied sequence of vectors. In the case of continuous speech, we would search for the sequence of models having the highest probability to have emitted this sequence of vectors.

In order to reduce the huge complexity flowing from an optimality exhaustive research, we can apply the *Token passing model* algorithm [10] which is in fact a particular implementation of the Viterbi's algorithm.

THE CORPUS

Sentences are continuously spoken by only one speaker.

The rough data which has been provided to us are the inner-lips height, width and area, synchronized with the acoustic signal [7].

This data set contains two-hundred files, each of which is composed of four letters from A to Z, continuously spoken in french, which will be separated into 70% of files for learning and 30% for test.

THE MODELS USED

We use one model for one word, which enable us to limit the problems of coarticulation and of the setting of boundaries. All the models have the same topology. While it is obvious that results suffer of that choice of topology, (a "A" should not have the same state number as a "W"), it makes the various comparisons and interpretations easier.

The labial model

Realization

The models used in this study are composed of eight states (six emitting states), each one having one transition to the following state and one loop (see Figure 1). Vectors have nine components: height, width, area, their speed and acceleration, divided up into three streams.

Results

We obtain a recognition rate of 42.05% on test data (data which are not learned). These results are surprisingly good, beyond expectation. It must be said on the point that the speaker is particularly cooperative.

The acoustic model

Realization

The chosen parameters are 12 cepstral coefficients spreading on a Mel's scale to which the energy of the signal as well as the derival of each of the former parameters are added.

The sources have the same structure as those used for the labial model, i.e. that all the words have the same topology.

Results

The model, when working on non-noisy data gives us a rate of 87.88% of word recognition.

The bimodal model

Realization

Cesptral coefficients being obtained every 10 ms and labial parameters every 20 ms, an interpolation of the latter is done. This interpolation (linear actually) is a cause of handicap for labial recognition.

The test we have done shows a loss of 10% of recognition between data sampled at 20ms and interpolated data. An alternative should be the use of Lagrange's or Newton's polynomial or splines, interpolations which have not been tested.

Acoustic and labial parameters described above are concatenated, and at last separated into two distinct streams, the topology of both acoustic and labial models is the same.

Results

The most part of information is contained in the acoustic source, so if this one is not noisy, labial information become some kind of noise. This explains the rate of 87.50% of word recognition.

NOISE INFLUENCE UPON RESULTS

We add a crowd noise (vocal frequencies) to the acoustic signal with a sound-noise ratio fixed. For this level of noise we do the learning and the test. Test results are calculated with the following formula :

$$R = \frac{N - I - S - D}{N} \times 100$$

N being a number of units to be recognized, I the number of insertions, S the number of substitutions, D the number of deletions et R the recognition rate.

sound-noise-ratio	labial model	acoustic model	bimodal model
no noise	42.05	87.88	87.50
6 dB	...	73.48	77.65
0 dB	...	53.03	62.88

CONCLUSION

The results we have obtained underline that with a non-noisy signal, labial movements carry some kind of noise, and make fall recognition rates.

However, results obtained in a noisy environment are encouraging and the applications of this type of system are numerous and are not limited to recognition : means, variances and probabilities calculated during the learning stage can be used for synthesis.

Recent studies have underlined that a lips display, in addition with vocal synthesis, considerably improves intelligibility [2].

FUTURE DIRECTIONS

The results we obtained are encouraging but could surely be improved. First, it is essential to increase our data base. It should enable us to have a more accurate learning and more significant tests.

We have to handle disynchronization between visual and acoustic information. It could lead to a modification of decoding algorithms as well as a modification of model structure.

At last, it is necessary to study further the management of weight tied to the various information sources according to various measures : sound environment, phonemes, acoustic or visual features, etc.

REFERENCES

1. C. Abry et M.T. Lallouache (1994) *Pour un modèle d'anticipation dépendant du locuteur - Données sur l'arrondissement en français* - Bulletin de la communication parlée

2. A. Adjoudani, T. Guiard-Marigny, B. Le Goff et C. Benoît (1994) *Un modèle 3D de lèvres parlantes* - 20èmes journées d'étude sur la parole

3. C. Benoît, T. Mohamadi, S.D. Kandel (1994) *Effects of phonetic context on audio-visual intelligibility of french* - Journal of Speech and Hearing Research

4. M-A. Cathiard et M.T. Lallouache (1992) *L'apport de la cinématique dans la perception visuelle de l'anticipation et de la rétention labiale* - 19èmes J.E.P. - Bruxelles

5.M. Gentil et L-J. Boë (1979) *Les lèvres et la parole : Données anatomiques et aspects physiologiques* - Travaux de l'institut de phonétique de Grenoble

6. M. Gentil (1981) *Etude de la perception de la parole : lecture labiale et sosies labiaux* - Etude pour le centre scientifique IBM de Paris

7. M.T. Lallouache (1991) *Un poste "visage-parole" couleur* - Thèse de doctorat - ICP Grenoble

8. E.D. Petajan (1884) *Automatic lipreading to enhance speech recognition* - Proceedings of the Global Communications Conference, IEEE Communication Society, Atlanta, Georgia, 265-272.

9.J. Robert-Rives (1995) *Modèles d'intégration audio-visuelle de signaux linguistiques : de la perception humaine à la reconnaissance automatique de voyelles* - Thèse de doctorat - ICP Grenoble

S.J. Young, N.H. Russel, J.H.S. Thornton (1989) *Token Passing : a Simple Conceptual Model for Connected Speech Recognition Systems* - Technical report - CUED/F-INFENG/TR.38

THE USE OF PROSODIC AGENTS IN A COOPERATIVE AUTOMATIC SPEECH RECOGNITION SYSTEM

Ph. Langlais and J.L. Cochard
IDIAP, CP 592, CH-1920 MARTIGNY
e-mail: langlais@idiap.ch cochard@idiap.ch

ABSTRACT

Two prosodic agents of a cooperative speech recognition system (namely $ETC_{vérif}$) will be presented in this paper. The first agent is processing information available in micro-prosodic variations. The second agent is dealing with linguistically-motivated aspects of prosody which are exceedingly useful to constraint solution space in a recognition task.

1. INTRODUCTION

It is often argued that prosodic can be used in numerous benefit ways in an automatic speech recognition process (ASR) [1]. Prosody is first involved in phonetically conditioned aspects (intrinsic values and coarticulation effects). It is well known for example that high vowels (such /*i*/) have an intrinsically lower duration than low vowels (such /*a*/). A recent study [2] measured a slight improvement of a large vocabulary speech recognition system by inserting a micro-prosodic model. Among all prosodic functions, one is of utmost importance from an ASR point of view: the grouping of related words in so-called prosodic words. Prosodic structuring can be useful for verifying and predicting linguistic organization proposed by other agents (syntactic or/and semantic ones). Numerous studies deal with this function and recent ones report interesting results for specific tasks such as disambiguation [3, 4]. Prosody is also reported to be useful in a speech understanding system specially in dialog situation where events such as repairs [5] and interrupts occur quite frequently [6]. This area is however far from the scope of this paper which will detail the first two enunciated points.

2. MICRO-PROSODIC AGENT

This agent is processing information available in intrinsic and co-intrinsic variations of fundamental frequency, intensity and duration parameters. It provides specially some weighted hypothesis to $ETC_{vérif}$ such as voice/voiceless diacritic recognition and voiced obstruent/non-obstruent consonant distinction. In this section we will just sum up special points that are described in depth, from a lexical access point of view, in [7].

Duration cues

We studied, on several corpora of French isolated words (ranging from 500 to 1000 words) uttered by several speakers, the vowels intrinsic durations and the right consonant effect (voice/voiceless and occlusive/constrictive) on the preceeding vowel. Durations were automatically obtained based on two different techniques; the first one is using the duration given by a lexical access module and the second one is based on non-contextual phonemic Markov models. We report hereafter major conclusions for this studies.

Even if high vowels durations are on average smaller than the ones of low vowels, intrinsic vowels durations are not reliable enough to be used in our system. In fact, only oral/nasal vowel distinction can be done (at least partially) with low error probability (see fig. 1).

Contextual effects can be observed on the average values but seem to be too fragile for classification techniques (error probability close to 0.4).

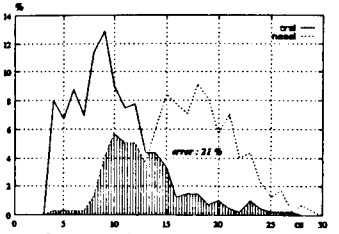


Figure 1. *Distributions of oral and nasal vowel durations on a corpus of 2 speakers' utterances of 800 tri-syllabic words.*

Intensity cues

We studied on the same corpora, the distributions of intensity values (measured by a classical raw power intensity) and we can conclude that local discrimination between vowels like /*a*/ and /*i*/ (see fig. 2) can be achieved with a reasonable probability error (at least for non final vowels), while pre-vocalic consonantic distinction can not.

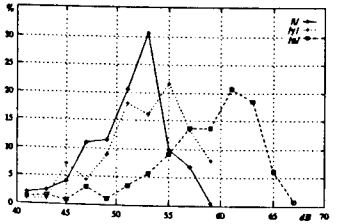


Figure 2. */i/, /y/ and /a/ distributions measured on initial vowels of tri-syllabic words*

Fundamental frequency cues

We measured this parameter with an algorithm implementing the *amdf* technique with satisfactory results. It provides a voice/voiceless decision based on the shape of *amdf* curve computed for each frame of signal. This method has been found very suitable for lexical filtering: more than 60% of a large lexicon was removed from potential candidates by the only mean of this decision with low error rate (less than 3%) on a task of 500 words recognition, each one uttered by 6 different speakers. In a top-down approach, the voice/voiceless distinction was useful too to re-rank lexical hypothesis with an average gain of 3 places.

As studied for duration and intensity, intrinsic and co-intrinsic frequency values have been considered and no robust information was discovered, except the obstruent/non-obstruent consonants distinction that can be achieved, at least partially, with reasonable error probability. The major cue of this distinction is the concave shape usually observed on non liquids intervocalic consonants (see fig. 3).

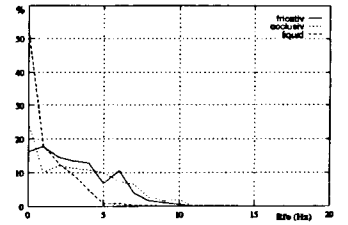


Figure 3. *Distributions of a concavity measure Rfo on different consonants classes.*

3. SUPRASEGMENTAL AGENT

Even if linguistically-motivated aspects of prosody gave rise to a lot of studies, there is not yet a unified model on which all researchers can agree. This is mainly due to the fact that prosodic phenomena depend on many distinct levels of linguistic representation. For example, even if it is well known that prosodic cues occur more frequently at syntactic constituents boundaries; this rule can be denied by other constraints such as rhythmic ones. Thus, learn-

ing by example seems to be an efficient way to solve this conflicting situation. We report here first results obtained by a suprasegmental agent based on this technique.

This agent makes use of an identification system of prosodic labels which points out, in a sentence, the occurrences of some particular prosodic cues (two-way emergence of a vowel fundamental frequency, lengthening of its duration, ...). The output of this treatment: a prosodic lattice, as well as the syntactic decomposition of the sentence, and its phonetic alignment (obtained by an automatic Viterbi alignment of allophones models) feed a statistical module which updates a knowledge source (KS). This KS quantifies — for a given corpus — the correlations between some syntactic, rhythmic and prosodic units.

Information is organized into a non-connected oriented graph. Each edge bears a syntactic and/or rhythmic constraint automatically derived from learning observations. Each vertex N (called a 'P-node') contains information such as the number of times it has been visited when processing learning data, the number of occurrences of each different prosodic label attached to observations and a tree structure $Crit(N)$ (called a 'SR-structure') describing the syntactic-rhythmic organization modeled by N. Each leaf node of the SR-structure holds all prosodic labels observed at the current constituent boundaries.

An observation O is a SR-structure with a fully described tree (*i.e.* syntactic tree with each node containing the right number of vowels). We denote p_o the depth of the tree and define d as the number of consecutive levels (beginning at root node) with fully instanciated numbers of vowels ($p \in [0, p_o]$ and $p \leq p_o$). $O_{p,d}$ is the SR-structure got from O by filtering out the p first levels with rhythmic constraints of the d first levels (*ex.* : $O_{p_o,0}$ is the syntactic structure of the observation, $O_{1,1}$ holds the number of vowels in the observation).

The graph grows automatically by updating each node holding a SR-structure that can be unified with $O_{p,d}$ and by creating missing nodes using the following algorithm:

$explore(N, O, p, d)$
  $if\ (p < p_o)$
    $if\ \exists N'$: P-node / $Crit(N') = O_{p+1,d}$
      $then\ update\ N'$
      $else\ create\ N'$: son of N
    $explore(N', O, p+1, j)$
  $if\ (d < p_o)$
    $if\ \exists N'$: P-node / $Crit(N') = O_{p,d}$
      $then\ update\ N'$
      $else\ create\ N'$: son of N
    $explore(N', O, p, d+1)$

An observation O can at most generate $\frac{p \times (p+3)}{2}$ P-nodes but in general factorization (depending on the application) significantly cuts down the graph expansion. This organization allows a user to easily query the system on particular syntactic-rhythmic structures, such as for example: NB(N1-999999(5).VIRG(2).N1-999(4)), that describes a number made up of three distinct groups with respectively 5, 2 and 4 vowels. The system answers by displaying figures of prosodic parameter contours (see fig. 4) modeled by the selected P-node with unifiable information and by providing a matrix describing the frequency of each prosodic label observed at this node.

This KS also provides a convenient way of scoring the adequacy between the measured prosodic cues and the syntactic-rhythmical structure that could be partially (internal node of the tree) or entirely (leaves of the description tree) defined in order to give weighted hypothesis for a specific input. We report hereafter (see fig. 5) results of a number recognition task. We feed the system with 500 numbers uttered by 70 speakers on a telephone line.

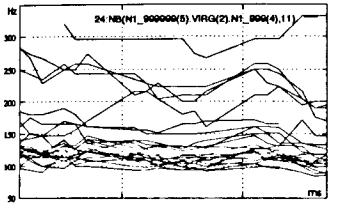


Figure 4. *Example of fo contours proposed by the system. Only initial, middle and final values of each group are taken into account and linearly smoothed.*

For each number, the system is provided with a syntactic structure, a phonetic alignment and a prosodic lattice (above 30 different labels). All these data are automatically computed from orthographic transcription. The system's objective is to predict the syntactic-rhythmic structure of a 100 numbers test data set. The score of a specific P-leaf is given by the maximum mark assigned to each possible path from the root to it. The score of a path is the average scoring of each of its P-nodes and a specific node in a path is scored by a distance between its local prosodic matrix and the input one. The figure 5 reports the ranking rate of the 100 observations.

4. DISCUSSION

This study demonstrates that most of intrinsic and co-intrinsic phenomena are difficult to handle, and only few cues seem to be useful for a recognition process. On this point, this study confirms Dumouchel's conclusions [2]. We propose a user-friendly and efficient system for scoring and/or predicting structural linguistic hypothesis that seems very promising for further investigation on prosody.

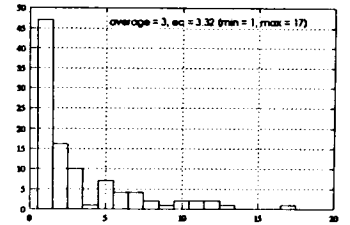


Figure 5. *Ranking of incoming candidates from an average number of possible ranks of 20.*

REFERENCES

[1] J. Vaissière. The use of prosodic parameters in automatic speech recognition. In *Recent Advances in Speech Understanding and Dialog Systems*. NATO ASI Series, 1988.

[2] P. Dumouchel. Suprasegmental features and continuous speech recognition. In *ICASSP*, pages 177–180, 1994.

[3] Andrew Hunt. A generalised model for utilising prosodic information in continuous speech recognition. In *ICASSP*, pages 169–172, 1994.

[4] Price and al. The use of prosody in syntactic disambiguation. In *DARPA workshop on Speech and Natural Language*, pages 372–377, Pacific Grove, Februar 1991.

[5] Nakatani and Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Acoustical Society of America*, 95(3):1603–1616, March 1994.

[6] Kompe and al. Prosody takes over: Towards a prosodically guided dialog system. In *Speech Communication*, pages 157–167, 1994.

[7] Béchet and al. Lexical filtrering by means of prosodic information. In *XIIIth ICPhS*, Stockholm, Sweden, 13-19 august 1995.

AUTOMATIC IDENTIFICATION OF ACCENTUAL-RHYTHMICAI STRUCTURE OF SPOKEN WORDS

B. Lobanov, T. Levkovskaya, E. Karnevskaya

Inst. of Egin. Cybernetics Ac. of Sc., Minsk, Belarus

ABSTRACT

A model of automatic identification of accentual-rhythmical structure (ARS) for isolated words is described. The identification of ARS is based on measuring of the duration, intensity and F1 value for each vowel of the word and their comparison with the set of patterns for all possible types of word ARS.

INTRODUCTION

The word ARS is generally defined as the distribution of force and distinctness of the articulation of the vowel sounds in the word. The acoustic correlate of the vowel articulation force is energy of speech signal, i.e. the intensity (amplitude) and duration, and the correlate of distinctness is stability of the vowel spectral characteristics.

It follows from the definition that words with a different position of the stressed vowel in a word display differences in ARS. The stressed vowel has the greatest force and distinctness in a word. If we assume that the stressed vowel articulation force and distinctness equal to 3 points, then the other positions of vowels in Russian words will be characteristics by the following values:

```
v v...v v V v...v v
2 1 1 1 2 3 1 1 1 2
```

It can be seen from the above scheme that the 2 points force is characteristic of the prestressed, initial and final vowels, while the force rest of the vowels is indicated with 1 point.

Information about ARS is a very important component of both the overall and phonemic recognition of a word [1,2]. In the overall recognition the knowledge of the word ARS makes it possible to considerably narrow the range of pretenders for the final decision about the word and thus increase the recognition reliability and speed , which is especially important in working with large vocabulary lists. In phonemic word recognition the knowledge of the degree of reduction of each vowel in the word, obtained as a result of the latter's ARS recognition allows to define more precisely the range of probable vowel allophones for each concrete position in a word and thus increase the probability of their correct choice.

EXPERIMENTAL STUDY

In essence ARS recognition is reduced to the identification of the stressed vowel position in a word. It is commonly known, that the stressed vowel possesses on average the maximal energy (product of duration and mean amplitude). However in real speech conditions this rule is not observed in all cases. As bas been shown by previous investigations, the energy value of a vowel is generally influenced by the following factors:
- the position of the vowel in relation to stress: prestressed, stressed, poststressed;
- the position in relation to word boundaries: initial, word-medial, final;
- whether the 1-st vowel in the word is covered or uncovered by a consonant;
- wether the last vowel in the word is open or closed by a consonant;
- wether the vowel being analyzed pertains to the narrow (u, i, ...) or broad (a, o, ...) articulatory group.

The aim of the present experimental study was to assess quantitatively the impact of the above factors on the vowel intensity and duration. The experimental corpus included Russian words containing from 1 to 7 syllables with a different position of word stress and different vowels in stressed and unstressed syllables. The selected words were recorded twice by two native speakers of the language and then digitized with the discretization at 16 kHz and 12 bits. The next step was marking the boundaries between phonemes by hand. Measurings of the vowel duration and intensity as well as the collecting of statistical data were carried out automatically. Besides, F1 frequency was found for each vowel so that a decision could be taken as to what group (narrow or broad) the vowel belongs. On the basis of the data obtained vowel duration and intensity distributions were received normalized with relation to the maximal value. These distributions covered words with different rhythmic structures and with different combinations of narrow and broad vowels: n-N-n, b-B-b, n-B-n, b-N-b, in prestressed, stressed and poststressed syllables. These data were then used for building up the common normalized distributions of the vowels duration, intensity and energy, corrected against the F1 meaning (fig.1).

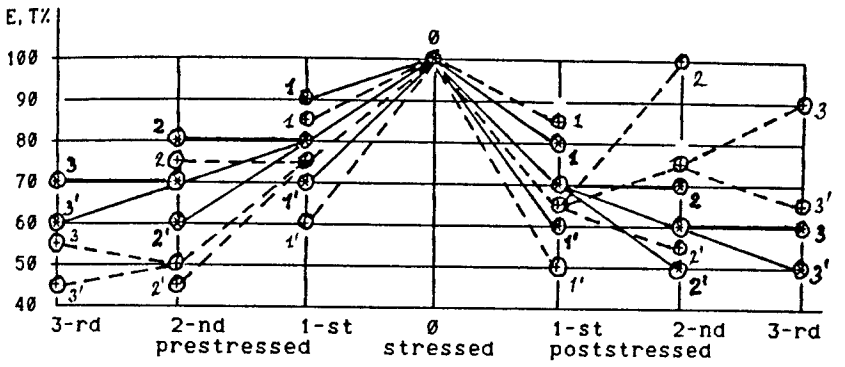


Fig.1. Experimental distributions of the corrected relative energyof vowels (solid lines - energy, dotted - duration).

In fig.1. the lines refer to the following ARS:
(1-0) - one uncovered prestressed syllable;
(1'-0) - one covered prestressed syllable;
(0-1) - one open poststressed syllable;
(0-1') - one closed poststressed syllable;
(2-0) - two prestressed syllables with an uncovered initial syllable;
(2'-0) - two prestressed syllables with a covered initial syllable;
(0-2) - two poststressed syllables with an open final syllable;
(0-2') - two poststressed syllables with a closed final syllable.
(3-0) - three prestressed syllables with an uncovered

initial syllable;
(3'-0) - three prestressed syllables with a covered initial syllable;
(0-3) - three poststressed syllables with an open final syllable;
(0-3') - three poststressed syllables with a closed final syllable.

If a word contains more than 3 prestressed or poststressed syllables, it is possible to continue using lines 3(3') - 0 - 3(3'), and in that case the energy value for the 2-nd prestressed vowel (or poststressed) is shifted to the 3-rd syllable if the total number of syllables is 4, or to the 3-rd and 4-th if the number of syllables is 5 or more, etc.

IDENTIFICATION PROCEDURE

The procedure of taking a decision about the ARS of the word being analysed is described in the following way. At the first stage a sequence of pretenders for the vowel phonemes in the word is identified. The phonemes are analyzed as to their belonging to the narrow or broad types according to whether the F1 frequency exceeds a defenite limit (in this case the limit was defined as F1 = 350 Hz). Next the energy value for each vowel is caiculated: Ei = Ti * Ai. This is then specified for broad vowels by multiplying by the coefficient Kn = 0,7, and for narrow vowels by the Kn = 1. Among the [E]i thus obtained is the maximal meaning, in relation to which the normalization of all energy values in the set is carried out. The normalized energies obtained are then compared with the set of pattern characteristics [En]i, similar to those in fig.1. The comparison is carried out with the following sample sequences. If the number of syllables is 2, an alternative comparison with one of the following pairs of sequences is made:

$$\begin{bmatrix} (1 - 0) , (0 - 1); \\ (1' - 0) , (0 - 1); \\ (1 - 0) , (0 - 1'); \\ (1' - 0) , (0 - 1'). \end{bmatrix} \quad (1)$$

The meaning (1') here is taken in the cases when the first vowel is preceded and the second vowel is followed not by a pause, but by one or more consonants. For a three syllable word comparison is drawn with one of the three-unit sequences:

$$\begin{bmatrix} (1 -1), (2 -0), (0 -2); \\ (1'-1), (0 -1), (0 -2); \\ (1 -1'), (2 -0), (0 -2'); \\ (1'-1'), (2'-0), (0 -2'). \end{bmatrix} \quad (2)$$

Similar sample sequences are composed for the number of syllables > 3. To take a decision about the type of the word ARS the energy similarity degree is calculated between each vowel in the word being analyzed and the set of energies pattern in the corresponding sequences (1), (2), ... Next the similarity sum is calculated for each of the possible sequences, shown in the brackets. Among the summed measures of similarity there is the one displaying the maximal degree. This sequence is assumed to be the most likely rhythmic structure of the analyzed word.

RESULT AND DISCUSSION

To check the effectivity of the suggested procedure a software model of the ARS automatic identification was worked out. The speech signal is analyzed with the help of a set of 5 octave filters in the range up to 8 KHZ in the time-window of 16 ms with a step of 8 ms. Out of 5 spectral parameters 3 segmenting parameters - P1, P2, P3 - are formed, displaying the maximal value of P1 - for the vowels, P2 - for the consonants and P3 - for the pauses. The segment boundaries are defined by analyzing the segmenting functions Sk obtained from Pk according to the formula

$$Skn = \sum_{i=1}^{L/2} (Sk, n+i - Sk, n-1)/L,$$

where n-is current timing, L-analysis window.

Within the boundaries determined in this way the vowels duration and intensity were defined, as well as the occurrence of a consonant or a pause before the first or after the last vowel of the word. In order to identify the vowels nature - high or low - estimation of F1 was carried out on the vowel segments by means of counting the number of zero-crossing at the output of the 1-st filter. The information obtained in this way was then used in the software of ARS automatic identification in accordance with the procedure described in the above section. The ARS automatic recognition algorithm was evaluated for its efficiency first on the speech material corpus described in section 2, and then on a new additional testing set of material. The results of the tests were summed up by fixing two types of errors:

Ms-the general percentage of erroneous identification of the word ARS;

Mr-the percentage of errors in the recognition of the ARS in the cases where the total number of vowels in the word and their articulatory type was identified and, besides, the decision about the presence / absence of a consonant at the beginning or at the end of the word was correct.

The results of both tests are estimated on average as: Ms=78%, Mr=94%. As is shown by the results of test evaluation the suggested procedure of the ARS automatic identification can be assessed as effective, on condition, however, that the preliminary phoneme segmentation and marking have been sufficiently correct. It should be noted hereby, that in certain cases, analyses for the degree of similarity (nearness) between the ARS being considered and the sample one, can help reveal the defects of phoneme segmentation and marking in the word. If some of the ARS have close similarity measures, they must be subject to further analysis with a view to making clear whether or not there is a vowel cluster or some mistake in segmentation.

REFERENCES

[1] Zue V. (1985), "The use of speech knowledge in automatic speech recognition", Proc. of the IEEE, vol. 73, N11, pp. 75-90.

[2] Carlson R. (1991), "Duration models in use", Proc. of the XII-th ICPhS, Vol 1, pp. 243-246.

HOW ACOUSTIC ANALYSES CAN IMPROVE SEGMENTATION CRITERIA

Christine Meunier

Laboratory of Psycholinguistic, 9 route de Drize, CH. 1227 Carouge, Switzerland

ABSTRACT

This paper shows how the acoustic analysis of transition phases between two phonetic units can improve the criteria of speech segmentation. We propose a system of rules based on acoustic analysis (of transition phase), phonetic distribution (type of cluster), and syntaxic context (isolated word, word juncture, etc). These rules allow us to validate segmentation criteria and to make them more precise.

1. INTRODUCTION

The elaboration of segmentation criteria and labelling conventions raises several types of problems [1]. One major problem is the lack of constancy in the segmentation criteria due to the of variability of speech. Indeed, we observe different transition phases between the same two consonants of a cluster (figures 2 and 3).

In addition, segmentation operations presuppose having prior criteria and, at the same time, those criteria can only be elaborated through the regular exercice of segmentation. A neophyte segmentator in this paradoxical situation experiences a short period of instability during which the criteria are established. Once they are stable, a human expert can describe the qualitative nature of the boundaries but he only has an approximative idea of the quantitative representation of boundaries.

Regarding the problem raised by criteria elaboration, we decided to divide the analysis into three different steps. First, basic criteria were drawn up; this means that only qualitative observations were made. Second, the analysis of the labelling data produced a quantitative representation of consonant transitions within clusters. Third, quantitative results are compared to qualitative observations.

2. METHODOLOGY

Some basic methodological options of our general study of French consonant clusters [2] are presented here to make this specific procedure clear.

2.1. Consonants

We classified the French consonants relating to articulation manner:
- "S" are Stops: p t k b d g
- "F" are Fricatives: f s S v z J
- "V" are Vocalic consonants: l r m n j w

Consonant clusters are the logical combinations of the 3 consonant classes:
- SS (as in "obtu")
- FF (as in "asphalte")
- VV (as in "parmi" or in "lui")
- SF (as in "psychologue")
- SV (as in "bleu")
- FV (as in "flou")
- FS (as in "style")
- VS (as in "halte")
- VF (as in "farce")

2.2. Segmentation principles

The segmentation we practiced is based on a methodology developped by Autesserre and Rossi [3]. It is an accurate segmentation and hierarchically organized labelling: phonetic units are segmented at three different levels: *macro-classes* (classes of phonemes), *phases* (onset, stable part and release of the phonetic units) and *attributes* (acoustic events). This segmentation procedure produces sub-segments which can be used to locate the boundary between two consonants. The acoustic analysis of these sub-segments will give us the statistical representation of the boundaries within clusters and also will help to quantify segmentation criteria.

2.3. Acoustic analysis principles

In segmenting and labelling a speech database, we were led to adopt a specific acoustic analysis methodology: we distinguished segmentation specific features (articulation manner and voicing) from coarticulation features (articulation place) [2].

An accurate analysis of the acoustic realisations of consonant clusters allows us to draw up a list of transition phases.

We define the sub-segment in the following way: it is a transitory segment which appears at and around the boundary between the two consonants. The transition phase between C1 and C2 is realised either directly (Direct Passage: the acoustic characteristics of C2 directly follow those of C1), or by a Transitory Segment which affects one of the two consonants. The transitory segment supposes a contrast to the acoustic characteristics of the consonants.

Consonant characteristics are described as follows:
-V: vocalic formant structure
-F: voiced or unvoiced noise
-S: silence or voicing + burst

Four types of transitory segments are considered:
- vocalisation: insertion of a vocalic or a vowel unit
- consonantisation: insertion of noise
- devoicing: disappearance of voicing
- voicing: appearance of unexpected voicing

2.4. Basic segmentation criteria: observation of the transition phase

A set of basic qualitative criteria are drawn up. They will be compared to the statistic analysis results. In order to give an impression of the nature of these criteria, we have presented below the transition phases for each cluster class.

DP is Direct Passage and TS is Transitory Segment:
-SS: DP or TS: devoicing
-FF: DP or TS: devoicing
-VV: DP
-SF: DP or TS: devoicing
-SV: if F voiced, DP
if F unvoiced, DP or TS: devoicing
-FV: if F voiced, DP
if F unvoiced, DP or TS: devoicing
-FS: DP or TS: devoicing
-VS: if F voiced, DP
if F unvoiced, DP or TS:
devoicing or consonantisation
-VF: if F voiced, DP
if F unvoiced, DP or TS:
devoicing or consonantisation

As we can observe, there are numerous types of transition phase between two consonants for each consonant cluster class. Only a quantitative analysis of the boundaries can make more precise and improve the segmentation criteria.

3. LINGUISTIC MATERIAL

It is important to base the acoustic analysis upon samples taken from different speech contexts. Indeed, we have observed that different speech types lead to specific types of transition phases within clusters. Therefore, our segmentation and analyses are based on three different speech contexts.

3.1. Isolated words

Our first segmentation criteria were elaborated from a corpus (Clusters: ACC01 to ACC05) of the BDSONS (the French sounds database [4]). This corpus is composed of 5 lists of 41 isolated mono or bi-syllabic words containing clusters. 12 speakers (6 male and 6 female) read them. The clusters were representative of their distribution in French.

3.2. Integrated words

The second corpus was based upon words integrated in the same sentence: "Ce n'est pas ____ qu'il faut dire"..

These sentences were read twice by two speakers.

3.3. Words in juncture

This corpus was made up of sentences in which clusters crossed word boundaries.These sentences were read twice by two speakers.

In this corpus we considered two levels of junctures: the first a major boundary and the second a minor boundary. In fact, the sentences were of the very simple syntactic structure: NP+VP. The first type of juncture was between NP and VP (the major syntactic boundary), the second one was inside VP (between V and N, the minor boundary). We expected to obtain different acoustic effects as a function of the type of juncture which separated the first and second consonant (C1 and C2).

4. RESULTS AND RULES

4.1. The distribution of clusters transition phases

The results of our acoustic analysis allows us to distinguish two different classes: clusters including a vocalic

consonant, and clusters without vocalic consonant. Vocalic consonants are systematically and regularly affected by assimilation effects [2] [5]. Acoustic variability obeys specific rules when a vocalic consonant is present in the cluster. Assimilation effects are also present in clusters without vocalic consonant, but the phenomena are quite irregular and do not follow specific rules.

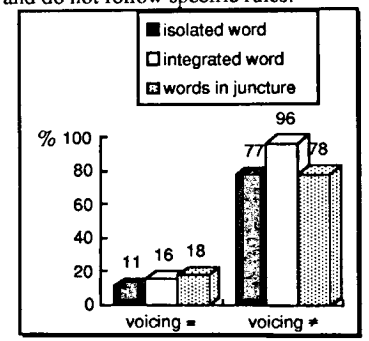


Figure 1: percentage of transitory segments within the three corpuses. Voicing similarity (voicing =) or opposition (voicing ≠) between the both consonants of the clusters are presented.

Voicing opposition is the factor which leads to the most important number of Transitory Segments (figure 1). Also the most frequent Transitory Segment is devoicing.

4.2. Rules

Rules are deduced from the acoustic analysis. The hierarchical presentation of rules follows the decreasing importance of variation in the transition phases of the clusters. The nature of the transition phase and their frequency are given (abbreviations are explained above).

1: If there is voicing opposition within cluster: 2 else TS0
2: If V is in cluster: 3 else 11
3: If cluster is pronounced in isolated word: 4 else 9
4: If V is C2: 5 else 8
5: If cluster is SV: 6 else 7
6: If V is /r/: TS1 else TS2
7: If V is /r/: TS1 else TS3
8: If V is /r/: TS4 else TS5
9: If V is C2: TS1 else 10
10: If V is /r/: TS6 else TS7
11: If C1 is voiced: 12 else 13
12: If there is a minor boundary between C1 and C2: TS8 else TS9
13: If cluster is FF: TS10 else 14
14: If there is a minor boundary between C1 and C2: TS11 else TS12

Lists and description of each Transitory Segment (TS):

TS0: irrelevant TS
TS1: TS = 100% of devoicing C2
TS2: TS > 90% of devoicing C2
TS3: TS > 60% of devoicing C2
TS4: TS > 5% of consonantisation C1
TS > 90% of devoicing C1
TS5: TS > 20% of consonantisation C1
TS > 15% of devoicing C1
TS6: TS > 30% of consonantisation C1
TS7: TS > 25% of consonantisation C1
TS > 20% of devoicing C1
TS8: TS = 100% of devoicing C1 (SS, FF, FS)
TS > 50% of devoicing C1 (SF)
TS9: TS = 100% of devoicing C1 (SS, FS)
TS > 50% of devoicing C1 (SF)
TS > 30% of devoicing C1 (FF)
TS > 50% of vocalisation C1(FF)
TS10: TS = 100% of devoicing C2
TS11: TS > 50% of devoicing C2
TS > 50% of voicing C1
TS12: TS > 80% of devoicing C2 (SS, FF)
TS > 30% of devoicing C2 (SF, FS)
TS > 30% of voicing C1 (SF, FS)
TS > 30% of vocalisation C1 (SF, FS)

5. DISCUSSION

These rules show that devoicing of a vocalic consonant in C2 position is the most regular phenomenon (TS1, TS2, TS3). This phenomenon becomes quite rare when the vocalic consonant is in C1 position (TS4, TS5, TS6, TS7), except for /r/ which is the most assimilated phonetic unit [2]. Tendencies are more uncertain for clusters composed of stops and/or fricatives. The devoicing of voiced C1 is frequent if C2 is unvoiced (TS8, TS9), but it depends on the cluster types.

Devoicing is the most important transitory segment. Nevertheless, the devoicing tends to be more progressive when a vocalic consonant is present in the cluster, and more regressive when clusters are made up of stops or fricatives.

5.1. Qualitative and quantitative segmentation criteria

These rules and, more precisely, the acoustic analysis of consonant clusters, confirm the basic segmentation criteria elaborated above. They bring quantitative precision for qualitative criteria. For SV and FV the devoicing is confirmed and represents the quasi systematic type of boundary; for VV, the Direct Passage is the most common boundary.; for FS devoicing is as frequent as voicing; etc.

Knowledge of the quantitative distribution of consonant cluster transition phases helps us to elaborate more robust segmentation criteria.

6. CONCLUSION

It is quite difficult to elaborate robust segmentation criteria. Nevertheless a human expert can formalize qualitative information and has a quite good representation of quantitative information. An accurate segmentation associated with an analysis of transition phases within clusters can make up for the lack of quantitative descriptions.

This additional segmentation tool should be useful for different speech technologies: Automatic Segmentation (improvement of reliability and precision), Speech Synthesis (improvement of clusters synthesis), Speech Recognition (improvement of segments and sub-segments identification).

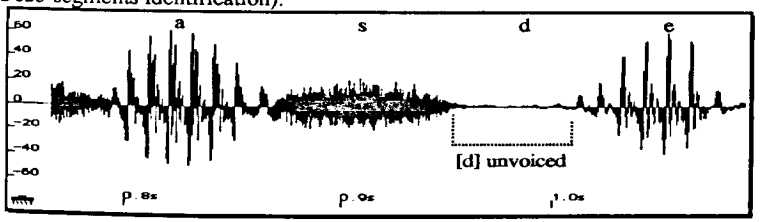


Figure 2: cluster /sd/ in the sentence "Les promeneurs chassent des papillons". The speaker produced a progressive assimilation: [d] is clearly unvoiced.

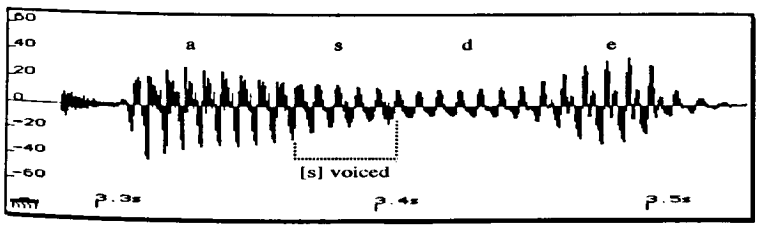


Figure 3: cluster /sd/ in the sentence "Les promeneurs chassent des papillons". The speaker produced a regressive assimilation: [s] is clearly voiced.

REFERENCES

[1] Abry, C., Autesserre, D., Barrerra, C., Benoît, C., Boë, L.J., Caelen, J., Caelen-Haumont, G., Rossi, M., Sock, R., Vigouroux, N. (1985), "Propositions pour la segmentation et l'étiquetage d'une base de données des sons du français", *Actes des 14èmes Journées d'Etude sur la Parole*, Paris, pp. 156-163.

[2] Meunier, C. (1994), *Les groupes de consonnes : problématique de la segmentation et variabilité acoustique*, Thèse de l'Université de Provence (Aix-Marseille I), Présentée le 7 mars 1994.

[3] Autesserre, D., Rossi, M. (1989), "Une méthode de segmentation et d'étiquetage des groupes consonantiques de la base de données des sons du français", *Journal d'Acoustique*, 2, pp. 311-322.

[4] Descout, R., Sérignat, J.F., Cervantes, O., Carré, R. (1986), "BDSONS: une base de données des sons du français", *12ème Congrès International d'Acoustique*, A4-7, Toronto.

[5] Meunier C. (1994) "Les facteurs responsables de l'assimilation: analyse de trois types de groupes de consonnes et de leur miroir", *Actes des 20èmes Journées d'Etude sur la Parole (JEP)*, Lannion, pp 447-452.

A MULTI-STAGE PROCEDURE FOR IDENTIFICAITON OF ACOUSTIC MICRO SEGMENTS IN STOP+VOWEL+STOP SYLLABLES.

Terrance. M. Nearey and Michael Kiefte
University of Alberta, Edmonton, Canada T6G 2E7

ABSTRACT

This paper describes an efficient multistage algorithm for the parsing of stop-vowel-stop syllables into acoustic microsegments typically used for scientific measurement of speech data and as well as in some ASR applications. The methods adopted illustate one way in which expert speech knowledge and powerful statistical pattern-recogniton can be usefully combined to provide robust and intuitively satisfying solutions.

OVERVIEW

Six microsegments of interest were established to characterize /CVk/ syllables, where the initial consonants ranged over /p, t, k, b, d, g/ and where the vowels ranged (in the test sets) over the 15 Canadian English vowels and diphthongs. The microsements are labeled *M1* to *M6* and are listed in Table 1.

Table 1. Microsegment categories for segmentation algorithm.

Microsegment	*description*
M1	initial (C) silence
M2	initial voicebar
M3	initial C release
M4	vowel
M5	final C silence
M6	final C burst

The object of the algorithims described is to produce reliable estimates for the beginings of the microsegments *M2*, *M3*, *M4* and *M5*.

The first stage uses statistics from a set of hand marked segments to provide a preliminary "fuzzy categorization" (in the form of a posteriori probabilities, henceforth APP scores) for each microsegment type and for each 20 ms rectangular window of speech advanced through the signal in 1 ms frames. Statistics are calculated from feature vectors for each frame which consist of six readily calculated properties. A second stage applies a Viterbi search method adapted from a continuously variable duration hidden (semi-)Markov model to bracket regions within which cursors can be placed. A third stage uses various heuristic measures to set the exact locations of specific cursors within the bracketed regions.

STAGE 1

The feature vectors for each frame consist of mean absolute amplitudes of 10 ms sections before and after the frame centers for 1) the original signal, 2) a high pass ($1/1-z^{-1}$) signal and 3) a band-pass ($1/1-z^{-2}$) filtered signal. The maximum mean absolute amplitude over all frames for the entire syllable was also included as a normalization measure.

Means, vectors and covarinace matrices were calculated using these features for the basic signal types shown in Table 2.

Table 2. Basic signal types defined by feature vectors.

Signal Type	*Description*
D1	silence
D2a	voice bar onset
D2b	voice bar cont.
D3a	C1 burst onset
D3b	C1 fric./asp.
D4a	V onset
D4b	V continuation
D4c	V end
D5 (=D1)	final /-k/ silence
D6a (=D3a)	final /-k/ burst
D6b (=D3b)	final /-k/ fric./asp.

The numbers associated with each distribution indicate the microsegment from Table 1 associated with each

distribution. Voice bar and consonant release are each associated with two distributions. Those labeled *a* (e.g. *D2a*) correspond to the onset part of the microsegment and those labeled with *b* correspond to continuation of the microsegment type. The distribution type "vowel" is associated with three distributions. Training samples for onset type *a* segments (as well as *D4c*) were centered on hand-marked cursors that delineated clearly acoustic boundaries, so that the features associated with the first and second 10 ms parts of the frame would be expected to differ significantly. Those for the continuation type *b* distibutions were more likely to be homogeneous. Training data for the *b* type distributions were extracted from the centers of hand-marked microsegment types of at least 30 ms duration to avoid the onset and offset of the microsegments) Only distributions *D1* through *D4a,b,c* were explicitly trained, since hand marked cursors did not exist for portions of the signal following *M4*. Distributions *D5* and *D6a,b* were effectively "tied" to the corresponding distributions for the variable initial stops.

APPs were calculated for distributions *D1* to *D4a,b,c*. Trials with a training set of 703 syllables indicated that these types could be identified with a reasonably high rate from the hand-marked data.

Running plots of the APP scores for stimuli were examined and indicated that the scores in question would serve as a useful basis for parsing the signal, since an appropriate distribution usually showed the highest APP for most of the duration of the target microsegments. (Appropriate distributions include either of the *a,b* pairs of microsegments associated with onset and continuation type distributions for reasons discussed below.)

STAGE 2

A modified continuously variable duration semi-HMM (CVDHMM) [1] was employed to group the frames of preliminary types *D1*-*D6*. The states of the CVDHMM are outlined in table 3 with their associated signal type. Possible transitions of the CVDHMM are shown in Figure 1.

The modifications from the CVDHMM of [1] consisted of three main simplifications involving "engineered" rather than optimized esteimates of parameter values. First, distribution parameters of Table 1 were fixed in advcance as described above and were not reestimated. Second, state transitions probabilities were determined *a priori*: the probability of the each exit path was set to the reciprocal of total number of such paths for that state. Third, although Gaussian state-duration distributions were originally investigated, it was found that they had little effect on the results and a simpler method ignoring state-durations entirely was adopted. This is equivalent to assuming a uniform distribution of all feasible durations (e.g. 1 to 1000 ms) for all states in a CVDHMM framework.

Table 3. CVDHMM states.

State	*Description*
S1 (D1)	silence
S2 (D2a)	C1 voicebar
S3 (D2b)	C1 voicebar cont.
S4 (D3a)	C1 burst
S5 (D3b)	C1 asp./fric.
S6 (D4a)	V onset
S7 (D4b)	V continuation
S8 (D4c)	V end
S9 (D5)	C2 silence
S10 (D6a)	C2 burst
S11 (D6b)	C2 asp./fric.

Preliminary observation of the results from the CVDHMM coarse segmentation showed that state changes were nearly always limited to points of large changes in the Stage 1 APP scores and that a substantial reduction of the search space for stage could be accomplished by using a simple preliminary thresholding of changes in APP scores. A measure of change in APP scores for each of the distribution types was calculated as

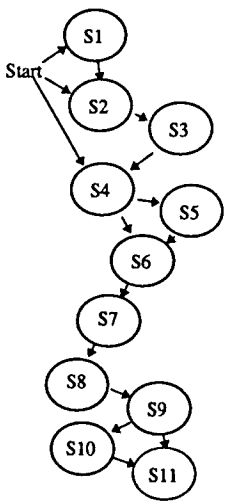
$$\delta_i = \sum_{k=1}^{n} \mathrm{abs}(\mathrm{W}(\mathrm{med}_3(m_{k,i-2}), \mathrm{med}_3(m_{k,i+1})))$$

where $m_{k,i}$ represents the six-point median filtered APP score for category k at time i, $\mathrm{med}_3(x)$ indicates the median of three arguments centered at point x and the function W is calculated as

$$W(x,y) = \frac{x}{1-\text{abs}(x-0.5)} - \frac{y}{1-\text{abs}(y-0.5)}$$

which weights the function in favor of high APPs. The threshhold was set at $\delta_i > 0.4$ where CVDHMM transitions were allowed to occur in the Viterbi search algorithm.

Figure 1. Non null state transitions for CVDHMM.



STAGE 3

Although the parsing provided by the CVDHMM was found to correctly bracket most of the events of interest, a final heuristic post-processing stage was used to fine-tune the placement of selected segment boundaries. The start of *M*2 was determined as

$$C_{vb} = \arg\max_t \left(\frac{L'(t,1)R(t,1)}{R'(t,1)L(t,1)+0.1} \right)$$

where R(*i*,1) is the average absolute value of the waveform in a 1 ms window to the right of point *i*, and L(*i*,1) is a similar measure taken from the left. R′(*i*,1) and L′(*i*,1) are the same measures taken from the low-passed filtered feature vector. *t* is limited to samples bounded by *S*2 and the first 30 ms of *S*3.

*M*4 was then located by exactly the same method with the values of *t* limited to the samples bounded by *S*6 and the first 30 ms of *S*7, with the exception that interval of averaging was 3 ms.

The following measure was then used to fine tune the start position of *M*3:

$$C_b = \arg\max_t \left(\frac{R(t,3)+R'(t,3)+A_{min}}{L(t,3)+L'(t,3)+A_{min}} \right)$$

where the integrated absolut value of the waveform is taken from 3 ms windows and A_{min} is the minimum mean absolute amplitude of the original signal. The search was restricted to the samples between the estimated start of *M*2 and the sample 30 ms following the start of *S*7.

The final position of *M*4 was taken as the window at which the product of the probability density function and the APP score was a maximum for *D*4*c* within the bounds of *S*8.

Informal evaluation of the algorithm indicates excellent agreement with human operator jugments in most cases. (More formal evaluations are in preparation and will be presented at the conference). The procedure has potential for use in semi-automated data collection for descriptive linguistic and speech database applications.

ACKNOWLEDGEMENTS

We would like to thank Tom Weltz and Terry Baxter for their programming skills. The research for this paper was supported by AGT.

REFERENCES

[1] Ljolje, A., & Levenson, S. E. (1991). "Development of an acoustic-phonetic hidden Markov model for continuous speech recognition", *I.E.E.E. Transactions on Signal Processing*, vol. 39, pp. 29-39.

Spotting prosodic boundaries in continuous speech in French

Vincent Pagel [1], Noëlle Carbonell [1], Yves Laprie [1], Jacqueline Vaissière [2]
(1) CRIN, BP 239, 54506 Vandoeuvre-Lès-Nancy, France
(2) ILPGA, 19 Rue des Bernardins, 75006 Paris, France

ABSTRACT

A radio speech corpus of 9mn has been prosodically marked by a phonetician expert, and non expert listeners. This corpus is large enough to train and test an automatic boundary spotting system, namely a time delay neural network fed with F0 values, vowels and pseudo-syllable durations. Results validate both prosodic marking and automatic spotting of prosodic events.

CONTEXT AND MOTIVATION

It is known for a number of languages that speech contains prosodic cues acting as boundary markers of different strength along the continuum. Boundary marking is particularly obvious in French, which has no distinctive lexical stress. Fundamental frequency (F0) movements are generally bounded by left and right word boundaries and phonemic lengthening marks the end of the sense groups. Besides, prominence is usually achieved through accents (F0 rises mostly) on monosyllables and on the first syllables of polysyllables. However, it is not clear whether and how prosodic cues may be used for segmenting continuous speech automatically.

Previous research using heuristic rules in expert systems [1][2], has uncovered problems, due mainly to: the diversity of intrinsic phonemic durations (nasal vowels are longer); the effects of the rate of speech (fewer and less obvious boundaries in rapid speech); inter-speaker variations; and the weighting of F0- vs. duration cues.

Moreover, in situations that favor expressiveness, accents may be misinterpreted as right word boundaries. This explains why current research on the automatic segmentation of speech into prosodic units applies to read speech only, namely to the exclusion of spontaneous oral communication where the expressive function of prosody prevails against its linguistic one.

We are currently studying «controlled speech», e.g. radio news announcements and press reviews, with a view to extending the scope of continuous speech recognition applications. The prosodic processing of «controlled speech» should prove easier than the analysis of spontaneous speech, since newscasters aim at and achieve balanced trade-offs between expressive and communicative purposes.

OBJECTIVES AND METHOD

The paper investigates the two following issues:

- Which acoustic parameters should be selected in order to discriminate left from right word/group boundaries accurately?
- Is the prosodic coding scheme we use consistent enough?

To answer these questions, we tested a multi-layer perceptron on a «controlled speech» corpus, using different sets of prosodic marks for the training stage.

Nine minutes of a radio press review, spoken by a single speaker, were phonetically labeled by a phonetician and prosodically coded both by a group of listeners and by an expert on prosody (J. Vaissière).

Twenty French phonetics students listened twice to the press review. They were asked to jot down (on the fly), first, the prosodic group end-boundaries they noticed (first audition), then the syllables they perceived as accented (second audition).

The expert coded F0 movements from a visual representation of the acoustic-phonetic data made up of: the phonetic segmentation marks and labels, the smoothed curve of F0, vocalic and intervocalic duration curves (all time-aligned and on the same sheet) computed from the phonetic segmentation. The wide-band spectrogram was also available but on a separate sheet.

ACOUSTIC-PROSODIC CODING

The expert described meaningful F0 movements and pauses, using a TOBI-like coding scheme we developed for French. Our coding symbols are presented in table 1; capital letters describe major F0 movements; indexes are used to indicate the position of the F0 movement inside the current word or prosodic group; symbols and indexes can be combined. For instance, the initial F0 rise in the current prosodic group is coded R- when it occurs on the first syllable of the current word; B+Rc indicates a crossing of the baseline followed by a continuation rise (at the end of the current prosodic group).

Results of the coding

The marks from 3 listeners were excluded from the analysis, since these subjects had difficulty in detecting accents.

The locations of the end-boundaries and accents perceived by the 17 remaining subjects were compared to both the expert's coding (F0 movement) and the computed vocalic and intervocalic durations, in order to:

-determine which prosodic phenomena characterize perceived word or group boundaries, and then identify efficient acoustic input data for the MLP;

- evaluate the size of the optimum context in terms of the number of sounds on the right and/or on the left of the marked syllable.

We shall not comment on the distribution and acoustic correlates of perceived word or group end-boundaries, because subjects' judgements agree with each other, and most syllables marked by them are followed by pauses which can be reliably detected by standard algorithms. Nevertheless, lengthening prevails over

Table 1. Our prosodic coding symbols

R : initial rise on the first syllable of a word	**Ri, Li** : movement delayed on the *ith* syllable
L : prominent fall on the last syllable of a word	**R-, L-** : on the head or tail of the current group

P : peak	**P^** symetric slopes	**P/** left slope deeper than the right one
	Ph particularly high F0 values	**P** right slope deeper than the left one

B crossing of the baseline
Rc continuation rise (last syllable of the prosodic group
S «sustained» (last syllable of prosodic groups)
U valley on a grammatical word (between two prosodic groups)
V sharp dip due to enhanced micromelody (separates two words)

pauses and F0 specific contours as a decision factor.

On the contrary, the distribution of perceived accents provides useful knowledge. 80 syllables were perceived as accented by more than 2 subjects (i.e. 1 perceived accent per 5,2 sec. time interval on the average). 10% of the perceived accents are syllables unmarked by the expert, half of them following a U mark; which suggests a rather limited influence of meaning on the perception of accents. The acoustic correlates for 5% of the marked syllables are atypical (for instance, F0 on the baseline), while the analysis of the other marks confirms previous studies: F0 (cf. Table 2) is the major cue for detecting prominence which may affect even grammatical words (16%); polysyllables are usually accented on the first syllable (85%), which is typical of news announcers' styles; lengthening is optional; when present (57%), it is moderate (compared to group end-boundaries) and affects consonants (64%) rather than vowels which may be shortened (10%).

Table 2. F0 movements corresponding to perceived accents

expert mark	mono-syllable	first syllable	last syllable
P	9%[a]	1%	5%
R-	11%[b]	11%	
R	15%	9%	
Total	35%[c]	26%	

a. 0% on grammatical words
b. 9% on grammatical words
c. Grand total reckon for 61% of the marks because other perceived marks where not coded by the expert

Besides, listeners' judgements favor unexpected phenomena against regularities: F0 peaks on the last syllables of polysyllables are generally ignored, as well as peaks on the second or third syllable; the first peak in a sequence of peaks (cf. digit sequences) is marked, while the following ones are not perceived as accents, even if they are more prominent than the first one.

These results indicate that local prosodic events provide useful reliable linguistic information on word and group boundaries, on condition that the interpretation of local phenomena involve contextual information on the long-term evolution of prosodic parameters

Prosodic segmentation using MLP

In the following, we use MLPs, implemented with cross-validation to avoid over-training, and with the softmax transfer function so that we get *maximum a posteriori probabilities* (MAPs) as described in [3]. When the training subset is not balanced, MAPs are divided by a priori probabilities for each class we want to recognize, so that we get scaled likelihoods. We decided that the system answers if one likelihood is greater than the sum of all the other likelihoods, so it is possible that the system gives no answer for a given input.
For each test, we use the last 75% of the speech corpus to train the MLP, and we perform the test on the first 25%.

We tried several inputs combinations as well as their derivatives: F0 average and regression coefficient on a vocalic segment, segmental duration, and pseudo-syllable duration. This last parameter is the time elapsed between the end of a vowel and the end of the next one, because in French the CV-CV syllable scheme is encountered most of the time. Note that we are not exactly in a true speech recognition situation, as the phonetic labeling gives vowels positions, but we do not consider it as a handicap since there exist reliable vocalic nucleus detectors nowadays.

Auditory marks

Two kinds of marks have been set by the listeners (*frontiers* and *accents*), which are attached to the syllable nucleus.
The MLP fed with any of the previously described values (F0,duration...), no matter the size of the temporal window, is not capable of reproducing the *accent* marking with a good score. Thus we consider that listeners' accent marks are not consistent, at least from a local point of view.

But for the *frontier* marks, the MLP fed with the duration, on a 5 vowel context, achieves the task with 11% insertion and 43% omissions.

Phonetician marks

At this stage, we use the auditory marks to select a significative subset of marks set by the expert. Considering the given number of mark types obtained, we found it necessary to gather them in generic classes to achieve a correct training of the MLP : R for initial rise (129 occurrences), P for peaks (128), B for baseline (105), C for continuation rise (50), Nil for no marking at all (1287).

Table 3. Confusion matrix: horizontally, expected results, vertically, MLP results. (356 answers / 400)

	Nil	B	C	P	R
Nil	227	6	0	4	3
B	7	20	0	0	0
C	0	5	7	1	1
P	5	0	1	25	5
R	0	0	0	5	34

After several tests, we kept vowel duration, F0 values, and pseudo-syllable duration on a 7 vocalic nucleus window to feed a MLP with 10 neurons in its hidden layer. The MLP has 5 outputs: one for each class mentioned above.

The MLP gives no answer for 44 configurations (concurrent answers).
Surprisingly, no nasality tag is required to draw the MLP attention on the fact that nasal vowels are much longer than vocalic ones.

RESULTS AND CONCLUSION

The main result is that this experience validates both the expert prosodic marking and the automatic spotting system. Furthermore, the confusion rate between P and R marks is rather low, which agrees with the results of [4]: lengthening is a more important correlate of F0 peak for P than for R. R marks recognized as P, are accented monosyllabics words.

The recognition rate for C is enhanced when we add F0 regression parameters, as involved vowels bear a long upward F0 move. However this adds a slight confusion in the identification of P marks.

Future work will aim at incorporating long term prosodic variations in the modelling of our prosodic marks.

REFERENCES

[1] J. Vaissière (1982), «A supra-segmental component in a French speech recognition system: reducing the number of lexical hypothesis and detecting the main boundaries» Recherches Acoustiques, Vol. VII, Lannion CNET, pp.109-125

[2] P. Langlais and H. Meloni (1993), «Integration of a Prosodic Component in an Automatic Speech Recognition System» in proceedings Eurospeech '93 Berlin

[3] Bourlard, H. (1993), *Connectionist Speech Recognition*, Kluwer Academic

[4] Pasdeloup, V. (1992), Durée intersyllabique dans le groupe accentuel en français, in proceedings XIX JEP, Bruxelles, pp. 531-536

A DESCRIPTIVE FRAMEWORK FOR THE INVESTIGATION OF SPATIO-TEMPORAL RELATIONSHIPS AMONG TRACK VARIABLES

*Nathalie Parlangeau *- Régine André-Obrecht *- Alain Marchal***
** Université Paul Sabatier Institut de Recherche en Informatique de Toulouse*
118, Route de Narbonne 31062 Toulouse Cedex
*** Université d'Aix en Provence I URA 261 Parole et Langage*
29, Ave R. Schuman 13621 Aix en Provence

ABSTRACT

Speech production is a complex process relying on coordinated gestures, but the acoustic signal does not depict its underlaying organization. Accepting that articulatory gestures are directly recognized through the coarticulation process, our proposal is to investigate the correlations between acoustic and articulatory informations in order to propose an intermediate level of representation and to assess gestural phonetic theory. We present here the framework of this investigation, the automatic labelling of the multi-sensor speech database ACCOR.

INTRODUCTION

To design Automatic Speech Recognition Systems, the main difficulty lies with the extremely large variability of the speech signal. This problem has been known and studied for a long time. One aspect is due to the assimilation and coarticulation phenomena : the assimilation is due to the phonological process whereas transitions between sounds are smoothed and phonetic features are spread over contiguous sounds. The coarticulation is inherent to the way speech is produced by the continuous motion of articulators [1]. Speech production is a complex process relying on coordinated gestures, but the acoustic signal does not immediatly reflect the underlaying organization. The question that arises is : what is the right level of representation ?

An hypothesis postulates that the articulatory gestures are directly recognized through the coarticulation process. From a theorical point of view, many researchers have seen in the articulation an intermediate level of representation which could link perception and production. The gestural phonetic theory is an alternative to previous theories like the motor theory which has been disproved as too simple [2]. Our proposal is to investigate the correlations between acoustic and articulatory informations in order to precise this intermediate level of representation, and to assess the gestural theory.

The first step of our study consists in automatically labelling the multi-sensor speech database which has been developed in the ESPRIT II Basic Research Action ACCOR (Articulatory Acoustic Correlations of Coarticulatory patterns) [3] . This database includes articulatory and aerodynamic as well as acoustic data. We dispose of five signals : the acoustic signal, the laryngograph trace, the nasal and oral airflow trace and the ElectroPalatoGraphic patterns.

METHODOLOGICAL FRAME

Considering that speech is the output of a production process which relies for its execution on coordinated gestures, the annotation should reflect articulatory timing ; what must be located are articulatory events and not segments [3]. The annotation of the database is based on the following two principles :

- **non-linearity**,
- **channel-independency** of information.

The first principle is adopted to lead to proper annotation and to not preclude any *a-priori* theorical assumptions about coarticulation. The methodological principle of channel-independency of the annotation is important to allow for the systematic investigation of the correlations between different levels of representation. We have added a third one which is the **robustness**, in the sense that each labelling method has to be speaker-independent and that the detections must be consistent.

All labelling methods are built on the same schema : we first detect the discontinuities on the signal, and we interpret them as indications of oncoming gestures from and towards articulatory goals. They are marked in the temporal domain according to precisely defined criteria.

THE ACOUSTIC SIGNAL

Articulatory goals

The labels currently used by phonetician experts, on the acoustic signal, are :

-VOW and VTW, Voice Onset and Voice Termination,

-SCW and SRW, Stop Closure and Stop Release of plosives /t/ and /k/.

For phoneticians, the label SCW means "a silence before a stop release" ; to preserve the non-linearity and to obtain systematic detections, we prefer to interpret this label as a simple closion before a silence.

Labelling methods

We first detect the acoustic discontinuities using a robust automatic segmentation method, the Forward-Backward divergence method [4] : the signal is assumed to be a sequence of stationary units, each one is characterized by an autoregressive model θ (L.P.C.). The method consists in performing on line a detection of changes of the parameter θ. The divergence test is based on the monitoring of a suitable statistic distance between two models θ_1 and θ_2. A change occurs when a threshold is exceeded. The procedure of detection is performed in parallel on the signal as on the high pass filtered signal. To avoid omissions, the signal is processed in the backwards when the delay between two boundaries is too long (100ms). The parameters (AR order, thresholds) are speaker independent.

Follows a first test to label segments as voiced/unvoiced/silence units. It is based on the mean variations of the energy, the correlation of the signal and the first reflection coefficient. The result is adapted using the zero level crossing ratio.

Next, we use a plosive detection test based on a Fourier Transform [5]. Two functions, the formantic energy Δ_n, and the high frequency energy variation Λ_n, are monitored. To detect a plosion, Δ_n must be lower than a threshold T1 and Λ_n must be higher than a threshold T2. We so locate voiced as well as unvoiced plosive bursts, the silence or the voiced segment before the burst.

Results and discussion

The events VOW and VTW are systematically found by our procedure, but we may observe a delay between the automatic position and the manual one. The table 1 gives an indication of these differences.

Large delays are specially present for the VTW event ; they are often due to a persistent sinusoidal wave which is present between the closure and the silence.

Table 1 : Number of automatic labels vs manual ones. Delay in ms.

	< 10	10<<20	>20
VOW	66/77	7/77	4/77
VTW	43/77	18/77	16/77

The SRW event corresponds to an unvoiced plosive burst. Our method detects the burst of all the phonemes /t/ and /k/, it detects also the labial plosive /p/ when it is located before anterior vowels.

THE LARYNGOGRAPH TRACE

Articulatory goals

Four articulatory events must be detected :

-VOX and VTX respectively Voice Onset and Voice Termination,

-PUX a Peack in an unvoiced segment,

-SGX a glottal stop.

Labelling methods

We use a simplified version of the Forward divergence method to detect the discontinuities of the laryngograph signal. Once the changes are detected, we interpret each segment as voiced/unvoiced using a voicing test based on an adaptative level crossing ratio which is applied for each segment on a centered window. We define two levels on both sides of the signal mean. We calculate the ratio between the two level crossing rates. VOX and VTX events are finally labelled according to very simple rules.

On the unvoiced segment, the event PUX squares with a change of gradient, so we make a regression interpolation. The PUX label results of a temporal coordination between the regression variations and distance from the VTX and VOX labels.

Results and discussion

Table 2 : Number of automatic labels vs manual ones. Delay in ms.

	< 10	10 << 20
VOX	26/27	1/27
VTX	26/27	1/27
PUX	15/24	

Good results are obtained from the VOX and VTX labels. For the PUX event, nine events are not found. These results are due to the lack of manual precise criteria ; this point is discussed with the experts. We observe some insertions due to the systematic application of the PUX rules.

The SGX event is not automatically detected because we have a single realization on the french sentences.

THE AERODYNAMIC AIRFLOW TRACE

The aerodynamic signals are the nasal and the oral volume velocity traces.

Articulatory goals

The nasal events to be detected are :

- BFN and DFN respectively Build up of airflow and Decline of airflow.
- MFN Maximum airflow.

The oral events are :

-BFN and DFN respectively Build up of airflow and Decline of airlow,

-MFN and mFN respectively Maximum and minimum airflow,

-SCO and SRO respectively stop Closure and Release.

Labelling methods

The recording technique for these signals is a pneumotach system using a Rothenberg mask. The drawback is the bad SNR of the signals. It is a problem concerning an automatic labelling, so we first filter the signals with a classic low pass band filter.

As articulatory events square with changes of gradient, we perform a regression interpolation. The application of specific rules gives us the final labelling for each signal.

Results and Discussion

Table 3 : Nasal airflow results. Number of automatic labels vs manual ones. Delay in ms.

	< 10	10<<20	> 20
BFN	42/65	7/65	8/65
MFN	3/12		
DFN	1/12	2/12	
MFDFN	43/49	1/49	2/49

We can see a seemingly bad result for DFN and MFN. In fact these two events are often labelled very closer, and our system detects an MFDFN event. Most omissions are explained by a too fine manual labelling.

Table 4 : Oral airflow results. Number of automatic labels vs manual ones. Delay in ms.

	< 10
MFDFO	49/49
mFDFO	24/32
SCO	13/22
SRO	24/28

Major problems occur when we have to detect the SCO event : SCO are generally confused with the mFO event and the criteria to avoid these substitions remain subjective.

THE EPG PATTERNS

Articulatory goals

The phonetician experts search events like Closure and Constriction and want to detect :

- for Closure,
  - ACE approach to closure,
  - SCE stop closure
  - MCE maximum closure
  - SRE stop release.
- for Constriction,
  - ACE approach to constriction
  - MCE maximum constriction
  - CRE constriction release.

Even if some events have the same labels, their detection depends on the context Closure or Constriction.

Labelling methods

As for the manual detection labelling, the automatic labelling is a dynamic process through the closure or constriction areas.

The patterns are 16*16 point images representing the tongue contact points.

For closure, we define three masks according to the three different closure configurations : a front , middle or back closure. The boundaries of the closure area precisely indicate the SCE and the SRE labels. The ACE is detected according to the place of the closure. It is a pattern in which there is a sufficient number of contacts around the center of the closure place. The MCE is the first pattern in the closure area, in which the number of contacts in the closure place is maximum.

For constriction, the method is now in progress ; we use the same approach to locate the constriction areas.

Results and Discussion

Table 5 : Number of automatic labels vs manual ones. Delay in ms.

	< 10	10<<20	< 20
ACE	58/77		18/77
SCE	78/78		
MCE	64/72	4/72	4/72
SRE	78/78		

The SCE and SRE detections are very robust. We almost indicate precisely the closure place.

The ACE label detection depends on the previous context. If it is a constriction, the detection has to be quite different. We do not take this difference into account, that explains delays greater than 20 ms.

To label MCE, different manual strategies are observed and we choose the frequent one. The delays are explained by this difference of strategies.

First experiments show correct detections of constriction areas.

CONCLUSION

We define an automatic labelling system for a multi-sensor speech database and quite good results are obtained. Discrepancies are due to the systematic nature of our procedures, and to the manual labelling criteria variations. This work permits to assess and to precise the manual labelling criteria.

Phoneticians are interested in these results for many reasons. First, the automatic labelling ensure the channel-independency of the annotations and it permits a robust application of defined criteria. The automatic procedure is also an important timesaver.

This work is the framework for the investigation of spatio-temporal correlations among track variables. It will permit to study an alternative to previous articulatory models for Automatic Speech Recognition [6].

REFERENCES

[1] J. VAISSIERE (1986)," Speech recognition : a tutorial ", ed. F. Fallside and W. A. Woods, Prentice Hall International, pp 191-236.

[2] A. M. LIBERMAN, I.G. MATTINGLY, " The motor Theory of Speech Perception Reversed ", *Cognition*, 21, pp 1-36.

[3] A. MARCHAL, W. J. HARDCASTLE (1993), " ACCOR : Instrumentation and database for cross-langage study of coarticulation", Langage and Speech, 2-3,pp 137-153,.

[4] A. MARCHAL, N. N'GUYAN-TRONG (1990), " Non Linearity and phonetic Segmentation", J. Acoust. Soc. Am., Suppl.1, Vol87, pp79-82.

[5] R. ANDRE-OBRECHT (1988), " A new approach for the automatic segmentation of continuous speech signals", *IEEE Trans on ASSP*.

[6] F. MALBOS, R. ANDRE-OBRECHT, M. BAUDRY (1994), " Comparaison de deux méthodes non paramétriques pour le détection des occlusives sourdes " , *XXèmes Journées d'Etudes sur la Parole*, pp175-180,.

[7] K. ERLER, L. DENG (1992), " HMM representation of quantized articulatory features for recognition of highly confusible words", *ICASSP 92*, Vol 1, pp 545-548.

FROM ACOUSTIC SIGNAL TO PHONETIC FEATURES: A DYNAMICALLY CONSTRAINED SELF-ORGANISING NEURAL NETWORK

Páll Steingrimsson[1], Bent Markussen[1], Ove Andersen[1], Paul Dalsgaard[1], William Barry[2]
[1]Center for PersonKommunikation (CPK), Aalborg University, Denmark
[2]Institute of Phonetics, University of the Saarland, Germany

ABSTRACT

Articulatorily based acoustic-phonetic features are derived from the speech signal via a Self-Organising Neural Network (SONN) using spectral and energy parameters calculated from single windowed segments of the signal, and dynamically constrained by a cost-minimisation procedure enforcing continuity on the basis of features present in the segment. Results of the smoothed feature traces are compared to a previously calculated, unconstrained feature output.

INTRODUCTION

The identification of the phonetic structure of an utterance in automatic speech recognition is seen increasingly as a hybrid task of combining pattern-recognition expertise with speech science knowledge [1-3]. Just as word recognition had to give way to recognition based on sub-word segmental units (phonemes or allophones) as the demand for ever larger vocabularies increased, so the segmental units have to give way to sub-segmental, parallel properties (features) as the realisation grows that, in normally produced (i.e. continuous) speech, the acoustic properties of a particular sound vary as a function of articulatory dependencies between consecutive segments [2,4,5]. The task of recovering the abstract phonemic structure from the acoustic signal is thus freed from the need to relate the overall acoustic pattern of a stretch of signal to a particular phoneme, and can exploit changes in particular features.

Two aspects of acoustic change are not usually differentiated explicitly:

(i) the fluctuation in spectral structure and consequent feature values during quasi-constant portions of the speech signal (fricatives, stop closures, nasals and laterals, and stressed-vowel centres), and

(ii) information-bearing spectral change (sonorant glides, diphthongs, and place-signalling CV- and VC-transitions.

Attempting to address the second area without dealing with the first is clearly inefficient. In addition, optimized feature extraction based on solutions to the first problem provides a potential basis for correcting for prosodic variation and vowel-to-vowel coarticulation.

This paper presents a method of employing dynamic constraints within a framework of a SONN for smoothing feature traces.

THE ACOUSTIC-PHONETIC FEATURE ESTIMATOR

The architecture of the feature estimator comprises two modules as shown in Figure 1.

The first, the preprocessing module, calculates, for each 10 ms windowed segment of the speech signal, a set of parameters consisting of 12 Mel-Frequency based cepstral, 12 delta cepstral, 12 delta-delta cepstral coefficients and the corresponding log energy, delta log energy and delta-delta log energy. The calculation takes place every 5 msec. From these parameters, a vector **a** is selected containing the 20 coefficients which maximise the phoneme separability.

The second module converts the vectors from the selected acoustic parameters **a** into acoustic-phonetic features ϕ by means of a SONN, the training of which is described in the next section (see [8] for further details).

The Self-Organising Neural Network

The SONN consists of a number of neurons - 400 are used in this research - which are arranged regularly in a 20 x 20 rectangular structure. Each neuron ***n*** has assigned to it a vector ***m***, the size and structure of which corresponds to the acoustic parameter vector

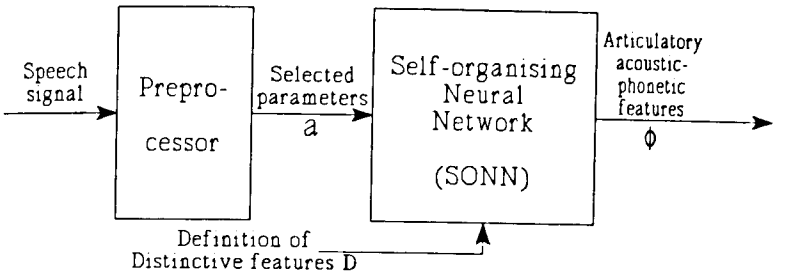


Figure 1. Architecture of phonetic feature estimator

a, and all neurons are connected in parallel to receive the same input.

The training session comprises three phases. Firstly, an unsupervised stimulation phase in which the SONN input is presented to speech data from the training speech corpus. Secondly, a supervised phoneme calibration phase and thirdly, a supervised acoustic phonetic feature calibration phase.

During the stimulation phase, each neuron n(x,y) of the SONN is assigned a parameter vector **m**(x,y), which is the weighted average of the acoustic vectors **a** firing (see e.g [8]) neuron n(x,y), during the entire stimulation phase.

As a result the SONN organises itself such that: 1) speech sounds that are acoustically close are represented in neighbouring neurons, 2) speech sounds which carry e.g. the same manner feature tend to group together in larger clusters, and 3) different speech sound classes, e.g. vowels vs consonants, are represented by neurons clustering in groups of classes.

Calibrating the SONN

I) The first calibration phase operates at phoneme level.

During the phoneme calibration phase, the SONN is submitted to the training speech corpus again, and the number of firings $n((x,y)|\phi_j)$ are registered for all phonemes ϕ_j and all neurons n(x,y), $x \in \{1 \cdots N_x\}$ and $y \in \{1 \cdots N_y\}$ within the SONN.

Given that n(x,y) is the neuron at position (x,y), $N_s(x,y)$, $s \in \{1 \cdots Q\}$, is a vector each element of which, $p((x,y)|\phi_j) = n((x,y)|\phi_j)/n(\phi_j)$, represents the frequency of occurrence that a specific phoneme is firing neuron n(x,y). The number of times $n((x,y)|\phi_j)$, that neuron n(x,y) is firing given that phoneme ϕ_j is present at the input, is calculated during the first calibration phase, and $n(\phi_j)$ is simply the total number of frames in the training corpus representing phoneme ϕ_j. By employing a clustering technique several vectors may be assigned to each neuron, i.e. $Q > 1$.

II) The second calibration phase operates at the level of phonological features.

Each phoneme ϕ_j is abstractly represented by a phonologically defined distinctive feature vector $\mathbf{D}_{\phi j}$, $j \in \{1 \cdots M\}$ where M is the number of distinctive features taken into account (observe that vectors **D** are dependent on the language). For example for the Danish phoneme symbol /u/, $\mathbf{D}_{/u/}$ is given by:

$\mathbf{D}_{/u/}$ = [[+voi] [+voc] [-fro] [-cen]
[+bac] [+rou] [+clo] [-mid] [-ope]
[-con] [0lab] [0den] [0alv] ...
[0fri] [0plo] [0sil]]^T

where '+' means feature *present*, '-' means *absent* and '0' means feature *not relevant.*

Based on vectors $\mathbf{N}_s(x,y)$ and $\mathbf{D}_{\phi j}$, $j \in \{1 \cdots M\}$, a phonetic framework vector $\mathbf{P}_s(x,y)$ is defined for each neuron n(x,y) [8]:

$$\mathbf{P}_s(x,y) \triangleq [\mathbf{D}_{\phi_1}\ \mathbf{D}_{\phi_2} \cdots \mathbf{D}_{\phi_j} \cdots \mathbf{D}_{\phi_M}]^T \times \mathbf{N}_s(x,y), \quad s \in \{1 \cdots Q\}$$

where Q is the number of phonetic framework vectors assigned to each neuron.

The elements $\mathbf{P}_s(x,y)(k)$ each represent an approximation to the probability that the k'th acoustic-phonetic feature has been involved in the firing of neuron n(x,y).

SONN FEATURE ESTIMATION

Previously the above expressions were used to estimate acoustic phonetic features directly from the acoustic speech signal on a frame-by-frame basis.

We have recently investigated new principles for estimating these features in which we include dynamic constraints in a Viterbi based minimisation of a chosen cost-function C(*l*) over a window extending back from the current speech frame *l*. The basic aim is to smooth the fluctuations in the feature values from one frame to the next.

The cost-function C(*l*) is chosen so as to contain elements which ensure that spectral changes as well as continuity of articulator movement are taken into consideration during the minimisation.

The first element is the summation of the distances $d_s[\bullet]$ between the incoming acoustic vectors **a**(l-i) and the neuron weight vectors **m**(x,y,l-i) as calculated over a window of fixed length L frames. This contribution is focused on the spectral differences within the window.

$$C(l) = \sum_{i=0}^{L-1} \left(d_s[\mathbf{a}(l-i),\mathbf{m}(x,y,l-i)] + w \cdot d_f[\mathbf{P}_s(x,y,l-i),\mathbf{P}_s(x,y,l-i-1)] \right)$$

The second summation adds a weighted contribution which is calculated on the basis the distances $d_f[\bullet]$ which represent the differences in the approximated probabilities given by the phonetic framework vectors $\mathbf{P}_s(x,y,l-i)$ and $\mathbf{P}_s(x,y,l-i-1)$ in the window. The factor w is a relative weighting between the two contributions.

Based on the minimisation, the resulting acoustic-phonetic feature vector ϕ is defined as follows:

$$\Phi(l-L+1) = \mathbf{P}_s(x,y,l-L+1).$$

ACOUSTIC-PHONETIC FEATURES

An example of a feature trace as estimated by the above procedure for Q = 2 is shown in Figure 2a on the next page.

The sentence 'pølsevognen stod midt' with the SAMPA transcription /0 p 2 l s @ v Q n s d0 d o D m e d0/ is transformed into phonetic features by applying the delineated approach.

A careful examination of the features illustrated in Figure 2a show a very close correspondence with the traditional definition of the phonemes as given in [8].

The feature traces shown in Figure 2a may be compared to the corresponding traces for the same speech signal as shown in Figure 2b, where the features are derived by the approach which performs the calculations on a frame-by-frame basis.

CONCLUSIONS AND OUTLOOK

The figures illustrate that articulatorily based features are indeed derivable, and that articulatory and functional features can operate together (see for example *VOC* and *VOI*, *VOI* capturing vocal fold activity, and *VOC* fairly successfully isolating vocalic segments). Also, as examination of Figure 2a indicates, the traces show a) acoustic dependencies between features that are used independently for phonological definition (see for example *BAC* and *ROU*), b) some clear changes in feature strenght during the time course of segments as defined by manual labelling (marked in figures, e.g. *OPE* for /Q/ and *MID, BAC, ROU* for /o/), and c) some carryover of features from the segment where a feature is relevant to where it is not (e.g. some vowel features into /l/ and /n/).

These are, at least in part, indications of articulatory transitions and coarticulation, which are not directly exploitable in a frame-by-frame system. The smoothed traces thus also provide a diagnostic base for the identification of phonetic events and features which require more dynamically oriented acoustic processing.

It is expected that the smoothed traces will provide a sounder basis for the estimation of segment boundaries and the identification of segments. Future work includes testing on two tasks which have been used previously to demonstrate the usability of the approach, namely that of automatic speech signal label alignment and that of phoneme recognition.

ACKNOWLEDGEMENTS

This work was partly funded via the support to CPK from the Danish Technical Research Council and partly supported by the Human Capital and Mobility Network project SPHERE (contract CHRX-CT93-0098).

REFERENCES

[1] Moore, R.K. (1993), " Whither a theory of speech pattern processing", Proceedings of European Conference on Speech Communication and Technology, pp 43-47.

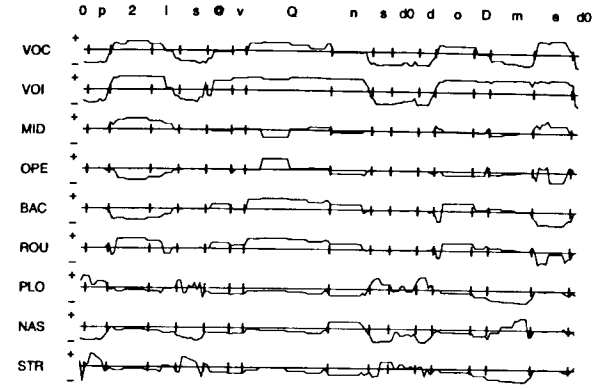


Figure 2a. Acoustic-phonetic features derived by the dynamically constrained approach

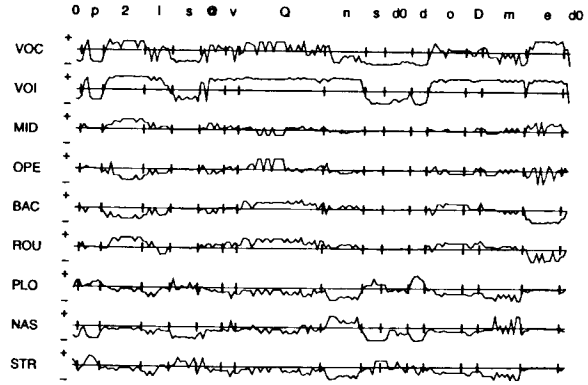


Figure 2b. Acoustic-phonetic features calculated on a frame-by-frame basis

[2] De Mori, R. & Flammia, G. (1993), "Speaker-independent consonant classification in continuous speech with distinctive features and neural networks", Journal of Acoustic Society of America 94(6), pp 3091-3103.

[3] Rose, R.C., Schroeter, J. & Sondhi, M.M. (1994), "An investigation of the potential role of speech production models in automatic speech recognition", International Conference on Spoken Language Processing 94, Yokohama, S12-1, pp 575-578.

[4] Deng, L. & Sun, D.X. (1994), "A statistical approach to automatic speech recognition using atomic speech units constructed from overlapping articulatory features", Journal of Acoustic Society of America 95(5), pp 2702-2719.

[5] Deng, L. & Sameti, H. (1994), "Automatic speech recognition using dynamically defined speech units", International Conference on Spoken Language Processing 94, Yokohama, S36-11, pp 2167-2170.

[6] Barry, W. & Dalsgaard, P. (1993), "Speech-database annotation. The importance of a multi-lingual approach", Proceedings of European Conference on Speech Communication and Technology, Berlin, pp 13-20.

[7] Dalsgaard, P., Andersen, O. & Barry, W. (1991), "The cross-language validity of acoustic-phonetic features in label alignment", Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-En-Provence, volume 5, pp 382-385.

[8] Dalsgaard, P. (1992), "Phoneme label alignment using acoustic-phonetic features and gaussian probability density functions", Computer Speech & Language, **6**, pp 303-329.

A NOISE-ROBUST SUBSPACE-BASED SOUND-CLASS DETECTOR

Wolfgang Wokurek
Institut für Maschinelle Sprachverarbeitung,
Universität Stuttgart, Germany

ABSTRACT

A computationally efficient approach to the automatic segmentation (labeling) of noise disturbed speech is presented. The segmentation algorithm employs short term spectrum based feature vectors and a subspace representation of the sound classes. The two sound classes of vowels and unvoiced fricatives are trained with the TIMIT acoustic phonetic continuous speech corpus. The sound class detector is applied in a speech enhancement system and for the automatic segment duration measurement.

INTRODUCTION

Originally this automatic sound class detection algorithm was developed as an improved replacement for the speech pause detector of a speech enhancement system [Wokurek 94]. Clearly this application requires noise robustness. Furthermore, a solution with low computational effort was sought to allow real time implementation. Representing the sound classes by subspaces meets both goals.

The speech signal is transformed into a sequence of feature vectors. This transformation controls which properties of the speech signal are represented by the length and by the direction of the feature vectors. In order to distinguish between different speech sounds the transformation will control the direction of the feature vectors by the shape of the spectrum. On the other hand, the feature vectors do not contain any pitch frequency information.

Unfortunately, no transformation is known that converts each phoneme or even each speech sound to a uniquely defined direction within the feature space [Furui 92]. Context, allophonic variations and noise ensure that the feature vector of every speech sound moves around quite a lot in the feature vector space. For automatic speech sound recognition it is necessary to describe the directions of the feature vector that are possible for each speech sound. A standard approach is to collect a number of representative feature vectors for each speech sound. Either the collection of feature vectors or a statistical description of them may be used as a representative of each speech sound. If the collection of feature vectors is used, their number is likely to exceed 1000. That number is met if e.g. every of 50 speech sounds is represented by only 20 vectors.

Is this large number of representative elements inevitably? A vector represents a single direction, in this sense it is a 'small' object in a vector space. A plane is a 'larger' object in that sense

— it contains infinitly many directions. Furthermore it only needs two (orthogonal) vectors to be defined. If even a plane is 'too small', D-dimensional subspaces could be used[1].

Is there any problem to represent speech sounds by planes instead of vectors? The problem might be that the plane is likely not only to contain the directions of the speech sound, but many other directions as well. So planes — or D-dimensional subspaces — should be used with care, and *not without further evidence.* In the case of this study it is observed experimentally that the feature vectors of the vowels [ieaou] lie in the vicinity of a plane. This motivates the notion of representing sets of speech sounds — sound classes — by subspaces.

Further experiments demonstrate the noise-robustness of a sound-class detector based on that subspace representation. These experiments indicate that the noise-robustness results from broadening the scope of discrimination from sounds to sound-classes. It should be noted however, that the sound-classes may not be defined arbitrarily. Only sounds with similar spectra are efficiently represented by a (low dimensional) subspace.

Finally it is important that the 'online' operation of this subspace-based sound-class detector is computationally efficient. Only the 'offline' training of the subspace representation of the sound-classes is computationally expensive.

FEATURE VECTORS

The disturbed speech signal is converted to feature vectors employing the short time energies of the output signals of a band-pass filter bank. Let each coordinate of the feature vector represent the short term energy within the bandwidth of a single channel of the filter bank. Then e.g. a formant produces a signal of high amplitude at the output of a certain filter bank channel and pulls the feature vector into it's direction. By this mechanism, the short term spectrum of the speech signal determines the direction of the feature vector. Each change in the formant structure changes the short term spectrum and turns the feature vector.

The short term spectrum of the speech signal is one ingredient to the feature vector direction — the bandwidth design of the filter bank is the second. Only those spectral changes turn the feature vector, that change it's coordinates; i.e. a formant movement turns the feature vector if and only if the formant moves into a different channel of the filter bank. Therefore different bandwidth designs of the filter bank are used.

The basis is the constant bandwidth of each channel. Here 20 channels with a bandwidth of 400 Hz are used to cover the frequency band from 0 to 8 kHz. A linear mapping occurs between the number of each channel and it's center frequency, therefore it is addressed as 'linear filter bank design'.

In contrast to that, a 'bark filter bank design' is employed to represent the psychoacoustic scale of critical frequency bands. Again, 20 channels cover the frequency band from 0 to 8 kHz, but the bandwidth starts with 100 Hz at low center frequency and increases to more than 1 kHz. Both filter bank designs are implemented using a computationally efficient algorithm — the 'Lerner filter bank' [Doblinger 91] [Lerner 64].

[1] A vector is a 1-dimensional subspace and a plane is a 2-dimensional subspace of the N-dimensional vector space.

The dimension of the feature vector space is defined by the number of filter bank channels. Hence the dimension is $N = 20$ for both, the linear and the bark filter bank design.

Given the noisy speech samples $x(n)$, the filter bank and the short term energy measurement of each channel results in the sequence of feature vectors $\mathbf{z}(m)$

$$x(n) \rightarrow \mathbf{z}(m) \quad .$$

The short term energies are computed with a time window duration of 20 ms. Therefore the feature vectors are low pass signals that do not require the sampling rate of the speech signal (20 kHz). The feature vectors are calculated at a sampling rate of 100 Hz, that is significantly lower. Hence the speech class detector operates at the lower sampling rate, what helps to limit the computational load.

SUBSPACE EXTRACTION

For the purpose of subspace extraction the feature vectors of all speech sounds that will train the considered sound class are collected within the observation Matrix

$$\mathbf{A} = (\ldots, \mathbf{z}(m), \ldots) \quad .$$

Now a subspace is required, that represents all these feature vectors in some sense. Here, the least square minimization of all vector components 'outside' (i.e. orthogonal to) the subspace is used as optimization criterion. A solution to that problem is found by the eigendecomposition of the correlation matrix of all feature vectors

$$\mathbf{C} = \mathbf{A}\mathbf{A}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^* \quad ,$$

where $\mathbf{U}$ is the orthogonal matrix of eigenvectors and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues [Golub 89].

Once the eigenvectors and eigenvalues of $\mathbf{C}$ are computed, the D dimensional subspace is spanned by the eigenvectors that correspond to the D largest eigenvalues. A threshold of 1% of the largest eigenvalue is used for automatic subspace dimension determination.

This algorithm represents each sound class and the noise signal by a single subspace. The automatic sound class detection is performed by comparing the actual feature vector to all subspaces. 'Winner' is the class that contains the largest component of the actual feature vector. Finally, a three point median filter improves the decision by removing isolated deviations.

SOUND CLASSES

Initially the two sound classes of vowels [i:e:a:o:u:] and unvoiced fricatives [fsçx$\int$] are trained with the TIMIT acoustic-phonetic continuous speech corpus. The subspace extraction results in a 1-dimensional noise subspace, a 2-dimensional subspace for the vowels and a 3-dimensional subspace for the fricatives. The resulting sound-class detector correctly detects the vowels, but confusions between the noise and the fricatives occur.

Due to larger spectral differences within the sound class of fricatives, the subspace of that class tends to 'catch' some of the feature vectors of noise segments. To minimize these errors, a further optimization of the sound classes is applied.

Now the sounds are exchanged between the classes until the smallest angle between every two subspaces is maximized. In addition, a smaller number of sound classes is preferred. The full search over all possible sound class partitions results in three sound classes. Due to better separation of the subspaces, the detection error rate decreases.

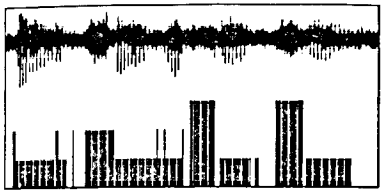


Figure 1: Detection of optimized sound classes: SNR=10dB, white noise

RESULTS

Figure 1 shows the segmentation of a noisy signal. The German utterance 'Deutscher Übersetzung' is analyzed. It is disturbed with white noise at a signal to noise ratio of 10 dB. The sound class detector employs the three optimized sound classes as well as the noise class.

Class #0 corresponds to the noise signal and is visible by 'missing' marks in Figure 1. Sound class #1 contains the voiced speech sounds. Sound class #2 is evoked by the $\int$, d and t sounds. Finally, the s sound is detected as a member of sound class #3.

The sound class detector is applied in a speech enhancement system to replace the speech pause detector. There, the speech spectrum estimation that is required for the enhancement, is controlled by the speech class decision. During noise segments an additional suppression is applied.

A second application of the automatic sound class detection algorithm is the automatic measurement of segment durations. The segment duration is used to normalize the time axis of fundamental frequency contours.

CONCLUSION

Classes of speech sounds are represented by low dimensional subspaces. The discrimination between sound classes instead of sounds is one source of the noise robustness of the algorithm; the restriction of sound class definition to sounds of similar spectral shape is the second. Finally, the subspace representation results in computational efficiency.

Even a small number of speech sound examples leads to reasonable results of the sound-class detector. This is due to the fact that a deterministic (nonstatistic) model is used. However, an increased number of different training sound examples improves the speaker independency of the segmentation results.

REFERENCES

[Doblinger 91] G. Doblinger. An efficient algorithm for uniform and nonuniform digital filter banks. In *IEEE Proceedings of ISCAS'91*, vol. 1, pp. 646 - 649. Singapore, 1991.

[Furui 92] S. Furui, M. M. Sondhi. *Advances in speech signal processing.* Marcel Dekker Inc., New York, 1992.

[Golub 89] G. H. Golub, C. F. Van Loan. *Matrix Computations.* The Johns Hopkins University Press, Baltimore, 1989.

[Lerner 64] R. M. Lerner. Band - pass filters with linear phase. In *Proceedings of the IEEE*, vol. 52, pp. 249 - 268. March 1964.

[Wokurek 94] W. Wokurek. *Sprachentstörung unter Verwendung eines Lautklassendetektors.* Dissertation, Technische Universität Wien. 1994.

SPEECH UNDERSTANDING DRIVEN BY CONCEPTUAL PROCESSING

Masao YOKOTA, Mariko ODA † and Seio ODA
Fukuoka Institute of Technology, Fukuoka, Japan
† Kurume Institute of Technology, Kurume, Japan

ABSTRACT

The authors have been constructing a speech understanding system IMAGES-S that can infer the conceptual information which the speaker would transmit. The processing for this purpose belongs no longer to wave signal processing but to natural language understanding, especially to conceptual processing with background knowledge such as commonsense, world-specific knowledge, etc. And moreover, understanding incompletely perceived speech is nearly equal to estimating the concepts of the words omitted in texts.

MODEL OF SPEECH UNDERSTANDING

Assume that one person "M_1" transmits his conceptual information "c" to the other person "M_2" acoustically in a language. The acoustic expression "r" of "c" which M_1 selects among the various paraphrases that he/she could generate is probably perceived by M_2 as a set of acoustic expressions "R_2" because of M_1's misstating or M_2's mishearing, or the noises during its propagation. Futhermore, each element of R_2 is interpreted as a set of conceptual information which in turn is merged into the total set "C_2", that is , the interpretation of R_2. These can be formalized as (1)-(3) below:

$$r \in \Phi_1(c) = R_1 = \{r_{11}, ..., r_{1l}\} \quad (1)$$

$$R_2 = \triangle_{12}(r) = \{r_{21}, ..., r_{2m}\} \quad (2)$$

$$C_2 = \cup_{i=1}\Phi_2^{-1}(r_{2i}) = \{c_{21}, ..., c_{2n}\} \quad (3)$$

where

Φ_i : M_i's acoustic verbalization process of conceptual information,

Φ_i^{-1} : M_i's interpretation process of acoustic expression,

and

$\triangle_{ij}$: the deformation process of acoustic expression in the environment of M_i and M_j.

The ideal speech recognition in M_2 will easily find "r" in R_2 because even the case $R_2 = \{r\}$ may happen. However, this is very difficult or almost impossible when the environment of the speaker M_1 and the hearer M_2 is not perfect, where "perfect" means "free from either mistakes or noises". Therefore, actually, M_2 is to select some "c'" among C_2 as would be "c" using background knowledge.

IMAGES-S simulates this process instead of the hearer M_2. That is, if the conceptual content "c'" resulting from understanding is reasonable, or not inconsistent with background knowledge, the system deems it as what the speaker would mean, and moreover, "r'", one of its verbalization "$\Phi_2(c')$", as what he would speak, where of course "r" is not always equal to "r'".

For an extreme example, IMAGES-S may transform such a dialogue between two persons as {"Where ?" " Bath."} into a more sophisticated one {"Where are you going ?" "I'm going to the public bath."}.

IMAGES-S consits of three modules: 1) Speech recognition (SRM), 2) Language understanding (LUM), and 3) Task realization (TRM). SRM transforms acoustic signal waves into word-lattices. LUM analyzes them syntactically and semantically and generates meaning representations, employing background knowledge. Finally, TRM realizes the tasks required by the speaker. Here is assumed that the task is limited to dictation.

CONCEPTUAL PROCESSING

LUM, utilizing the background knowledge K_B, estimates the concepts of the words unrecognized in SRM and such an inference process can be formalized as (4).

$$I(P[x_1, ..., x_n] \wedge K_B \vdash I(P[p_1, ..., p_n]) \quad (4)$$

where

$P[.]$: imcompletely recognized speech,

x_i : word-sequence not recognized or recognized with a very low likelihood,

and

p_i : estimated word-sequence.

The inference process succeeds when $I(P[.])$ is unified with background knowledge K_B, which superficially, results in substitutions θ_s in (5).

$$I(p) \wedge K_B \vdash I(P\theta), \theta = \{x_1/p_1, ..., x_n/p_n\} \quad (5)$$

The total process is formalized as (6)-(8) below:

$$h(P, K_B) = H \quad (6)$$

$$H = \{P[x_1, ..., x_n]\theta | hypothrtical\ restored\ word\text{-}sequence\} = \{H_1, ..., H_m\} \quad (7)$$

$$e(H) = H' \quad (8)$$

where

h : hypothsis generating function,

H : a set of hypotheses,

e : adequacy evaluating function,

and

H' : ordered H according to a certain preference.

At present, the prefernce order is determined according to the hypothesis as follows : "What is most easily understandable is the best understanding result.". This determination is realized by calculating the complexity of understanding. The representation of knowledge or speech contents in our system is based on the first-order predicate logic and the complexity is deemed as the total cost (C_t) of translation from a surface structure(i.e. sentence) into a conceptual structure(i.e. logical formula). The authors have found C_t given nearly by the equation (9) which approximates the total times of variable unification, predicate insertion, etc. occurring through the translation process.

$$C_t = 2N_0 + W + E \quad (9)$$

where

N_0 : the number of the words recognized in SRM ,

W : the total number of the words inferred in LUM,

and

E : the number of the words representing objects or events inferred in LUM.

The understanding result with the least C_t is determined as the best.

Assume that EX-7 below is one of the word-sequences generated from the word-lattice put out by SRM. LUM ,using the background knowledge in Table 1, understands it and TRM generates such sentences (S1.1)-(S2.5). The underlined parts are the inferred and inserted words.

The preference order among the sentences is shown by P in Table 2, which implies that S1.1 is the best and that S2.3 and S2.4 are the worst.

EX-7 父親·X_1·自動車·X_2·学校·X_3·通勤.
(= Father·X_1·automobile·X_2·school·X_3·commute)

S1.1 父親 が自動車学校に通勤.
(= Father commutes to the automobile school.)

S1.2 父親 が経営する 自動車学校に誰かが通勤.
(= Someone commutes to the automobile school owned by Father.)

S2.1 父親が自動車で学校に通勤.
(= Father commutes to the school by automobile.)

S2.2 父親の所有する自動車で学校に誰かが 通勤.
(= Someone commutes to the school by the automobile owned by Father.)

S2.3 父親の経営する自動車について教育する学校に誰かが通勤.
(= Someone commutes to the school for automobile education which is owned by Father.)

S2.4 父親の所有する自動車のある学校に誰かが通勤.
(= Someone commutes to the school where the automobile owned by Father is.)

S2.5 父親が自動車について教育する学校に通勤.
(= Father commutes to the school for automobile education.)

CONCLUSION

The modules LUM and TRM are almost equal to IMAGES-II[1, 2], that is, almost completed. The simulation of these modules has proved the validity of IMAGES-S. In near future, C_t will be improved in order to reflect coherence and cohesion in context. The problem left unsolved is the connection of SRM and LUM. The module SRM will be realized by employing Hidden Markov Models (HMMs).

References

[1] Yokota,M. et al (1984), Language-picture question-answering through common semantic representation and its application to the world of weather report, in (Bolc,L. ed.) Natural Language Communication..., Springer-Verlag

[2] Yokota,M. et al (1991), "On Natural language understanding system, IMAGES-II", IEICE Japan

[3] Flanagan,J.L. (1994), Technologies for multimedia communications, Proc. of IEEE, 82-4

Table 1: A part of background Knowledge (word-meanings)*

word	word-meaning = [concept : unifying operations :]
通勤	通勤$(x) \Leftrightarrow L(z, y_0, y_0, y_1, A_p) \wedge$... $\wedge$通勤$^{+}(x) \wedge$人間$(y_0) \wedge$施設$(y_1) \wedge$ 手段(z):ARG(dep(が),y_0),ARG(dep(に),y_1),ARG(dep(で),z),...;
学校	学校$(x) \Leftrightarrow$施設$(x) \wedge$教育$^{++}(y, x, ...) \wedge$...: ;
父親	父親$(x) \Leftrightarrow$男$(x) \wedge$親(x): :
自動車	自動車$(x) \Leftrightarrow L(\phi, x, p, q, A_p) \sqcap L(x, y, p, q, A_p) \wedge p \neq q \wedge$ 自動車$^{+}(x) \wedge$ 物$(y) \wedge$所有$(z_1) \wedge$教育$^{++}(z_2, ..., x, ...) \wedge$...: ;
教育	教育$(x) \Leftrightarrow$教育$^{+}(x) \wedge$教育$^{++}(x, y, z, ...) \wedge$施設$(y) \wedge$事物$(z)$; ARG(dep(が),y),ARG(dep(について),z),... :

*$\sqcap$: "simultaneously AND", $L(X, Y, U, V, A_p)$: "Y moves from U to V by X", ϕ : "don't care", 男 : "male", 親 : "parent", 施設 : "institution", 物 : "object", 事物 : "object or event", ARG(X, Y) : "unify X with Y".

Table 2: Evaluation of understanding results *

I.D.	θ	N_0	W	E	C_T	P
S1.1	{X_1/が, X_2/ϵ,X_3/に}	3	2	0	8	1
S1.2	{X_1/が経営する, X_2/ϵ,X_3/に誰かが}	3	5	2	13	3
S2.1	{X_1/が, X_2/で, X_3/に}	4	3	0	11	2
S2.2	{X_1/の所有する, X_2/で, X_3/に誰かが}	4	6	2	16	5
S2.3	{X_1/の経営する, X_2/について教育する, X_3/に誰かが}	4	7	3	18	6
S2.4	{X_1/の所有する, X_2/のある, X_3/に誰かが}	4	7	3	18	6
S2.5	{X_1/が, X_2/について教育する, X_3/に}	4	4	1	13	3

*ϵ means empty.

AUDITORY ORGANIZATION OF PROMINENCE AND CHUNKING IN SPOKEN SWEDISH

Robert Bannert
Department of Phonetics, Umeå University, Sweden

ABSTRACT

In an investigation of the macro-prosodic organization of spoken Swedish, different aspects of the listeners' variation were studied. Two listener groups, students at the beginner's level and trained phoneticians, had to mark the most prominent words and the chunks they could hear in speech samples of spontaneous Standard Swedish. Variations concerning each prominent word and chunk, the number of scores per item and the number of scores per listener are presented.

INTRODUCTION

In the past, prosodic organization of speech has been studied mostly in texts read-aloud. In order to arrive at a theory of speech, it is imperative to investigate the macro-prosodic structure of spontaneous speech from a perceptual point of view. A research programme with this goal was launched some time ago [1, 2], continuing previous research [3, 4]. Two significant prosodic features were selected, the highest degree of prominence (focus accent) and chunking (phrasing).

For this pilot investigation, aiming at the development of methodological insights into the study of perceptual modelling of the macro-prosodic organization of spoken Swedish, two samples of spontaneous (monologue) speech were used. They were produced by a female and a male speaker of Standard Swedish, about 40 years old, both with academic backgrounds. The speech sample of the female speaker contained 196 words, thus, there is a possible 196 possible votes for prominence and 195 votes for chunking . The speech sample of the male speaker contained 186 words and 185 possible chunk boundaries. Each speech sample had the duration of approximately one minute. The listeners had to mark the most prominent words and the chunking on a sheet of paper where the speech samples were given in orthographic representation. However, no punctuation marks were used. Four listener groups participated. A group of 26 students and of 5 trained phoneticians scored separately for prominence, another group of 29 students and 3 experts marked for chunking. The speech samples were presented from a loudspeaker four times.

Table 1. General distribution of the listeners' scores: prominence, chunking; students, experts; three categories. The first line gives the number of scores, the second line the percentage.

	Prominence					
	female speaker			**male speaker**		
scores	0	≤ 50%	> 75%	0	≤ 50%	> 75%
26 students	107 55	68 35	9 6	87 47	83 45	3 2
5 experts	143 73	34 17	13 7	142 76	23 12	15 8

	Chunking					
	female speaker			**male speaker**		
scores	0	≤ 50%	> 75%	0	≤ 50%	< 75%
29 students	135 69	46 24	6 3	125 68	44 24	8 4
3 experts	155 79	20 10	13 7	151 82	17 9	13 7

LISTENERS' SCORES

First the general distribution of the scores will be given, followed by the marks for prominence and chunking.

General distribution

As a first rough measure of the distribution of votes (markings, scores) of the listeners, a simplified account of the data is given in Table 1. The complete data is to be found in [5].

From Table 1 it can be seen clearly that experts vote in a much more consistent way across all the categories and speakers than the students. This difference could be expected, although the instructions for both groups were formulated in an identical, though general way of expression. The experts, it has to be assumed, reformulated and defined the instructions in phonetic-prosodic terms which the students had not learned yet.

Prominence and chunking

The group scores for each word in the first part of the speech sample of the female speaker are shown, prominence in Figure 1, chunking in Figure 2. The results are representative for the rest of this speech sample and also for the male speaker's speech sample.

The histograms of Figure 1 show a rather large variation in the scoring of the listeners and listener groups. In some words, students and experts agree rather well and to a high degree, in other words they score quite differently.

Figure 2 shows the percentage of the chunking for the students and the experts. Even in this case, listeners vary considerably in hearing chunks.

VOTES PER WORD AND CHUNK

Instances of numbers of scores per word for prominence for the students are given in Figure 3 for the female and male speech samples. No word received the highest number of possible votes per word, namely 26. Contrary to the experts, the students show a rather even and low distribution over the whole range, except for the lowest part, 1 and 2 votes per word.

Even this observation can be interpreted in the same way as above, namely that experts are more consistent in their scores due to their knowledge of prosody and the acoustic correlates of

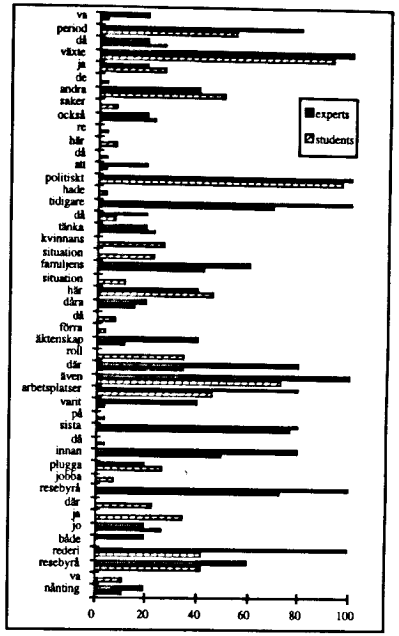


Figure 1. Percentage of total scores for the most prominent words by 26 students and 5 experts. Female speaker, first part of speech sample.

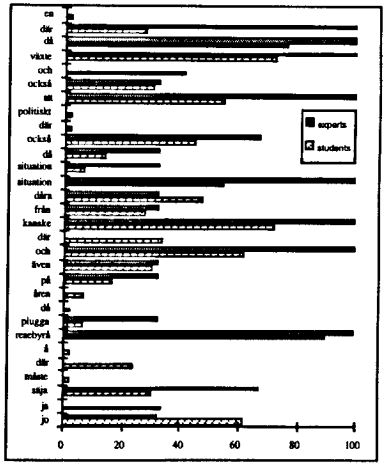


Figure 2. Percentage of the segmentation of the speech sample into chunks by 29 students and 3 experts. Female speaker, first part of speech sample.

prominence. However, it should be noted in any case, that experts, too, show

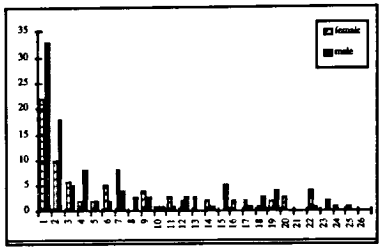


Figure 3. Instances of number of scores per word (prominence). Female and male speaker.

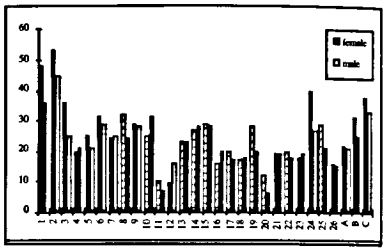


Figure 4. Individual distribution of listener scores for prominence. Students and experts, female and male speaker.

a relatively high degree of uncertainty, expressed in the rather high number of only one vote per word, especially with respect to the female speaker. For this bias, there does not seem to be an easy explanation because three of the experts were woman.

The scores for chunking appear to be similar for students and experts, i.e. both groups show a high degree of uncertainty. The student group gives the highest numbers of votes, 28 and 29 respectively, only to 5 chunks for the male speaker and only to 1 chunk for the female speaker. There were 40 chunks that received 1-28 votes (cf. Table 1). The experts give three votes, the highest number for this group, to 15 chunks out of 40 for the female speaker, and to 13 chunks out of 34 for the male speaker. Only one vote per chunk is given to 16 chunks for the female and to 14 chunks for the male speaker.

In comparison, the data suggest that students have great difficulties in recognizing the highest degree of prominence (focus accent) and chunking in spontaneous Swedish. Experts do better in recognizing focus accent. They, too, are rather bad at assigning chunk boundaries unanimously.

LISTENERS' VOTES

As an illustration of the individual variation among listeners, Figure 4 shows the scores for prominence for both speech samples and both groups. It is striking to observe how great the difference between raters can be. The lowest number of votes for prominent words in the texts, namely 6 votes, is given by listener no. 20 for the speech sample of the male speaker. This listener and listener no. 11 (7 votes) are very thrifty when they spend their votes. At the opposite end of the range we find listener no. 2 who gives 53 votes for the most prominent words of the female speaker. This means that one word out of four is heard as a most prominent word by listener no. 2, while listeners nos. 20 and 11 only hear one word out of 30 as most prominent.

The most striking aspect, even for chunking, is the great variation between the individuals. The average score for the students is 19.4 votes for both speakers. The experts' average score is almost 25 votes for both speakers.

AGREEMENT BETWEEN STUDENTS AND EXPERTS

In spite of all the inter- and intra-group variation, accounted for in this paper, there is one clear difference to be noted between the groups. In Table 1, it could be seen that experts, trained phoneticians specializing in prosody, score more consistently across varying conditions compared to students who have no training in prosodic theory and labelling, nor experience in carrying out such a listening test. An interesting question arises: How much agreement is to be found between the students and the experts in identifying prominence and chunking in spontaneous speech?

In order to give a quantitative answer to this question a statistical analysis, a simple regression analysis, was conducted, given the assumption of a linear relationship between the variables and their independency.

Only in one case, the scores for chunking of the male speaker, is the regression coefficient high, 0.97. In the other three cases, it is about 0.7. It can be interpreted such that there is not an excellent, although a rather good agreement between the scoring of the students and the experts. However, when the coefficient of determination is taken into consideration, this interpretation has perhaps to be restricted. Although about two thirds of the variation is explained by the independent variable (0.647 - 0.688), in one case not even half of it is explained (0.453). Therefore, the two listener groups do not agree well in scoring prominence and chunking in spontaneous Swedish.

CONCLUSIONS

When the pioneering, and now classical, work by Gårding [7] was published, prosodic research at then time, it could be said, was in its infancy. Since then significant contributions to the understanding of prosody and its role in speech communication have been made. One great insight in the dimension of stress or prominence in Swedish was achieved by Bruce [6] where he demonstrated convincingly that the famous and puzzling Swedish word accents have to be isolated from focus or phrase accent. Focus accent that signalled the most prominent words depending on context is mainly characterized by a tonal rise following the word accent fall in Standard Swedish. This separate tonal rise is a very marked cue and is easy to be heard. Therefore it was expected that listeners would easily hear focussed word but would have difficulties to decide upon chunking.

At a first glance, the results of this study appear to point to the interpretation that focus accent and phrase boundaries are non-existent in spontaneous speech or that listeners organize spontaneous speech in quite different ways using maybe divergent strategies. However, there are strong reasons to believe in the opposite interpretation. Listeners process the speech flow by applying rather general macro-prosodic strategies. This does not mean of course that listeners would identify prominent words or chunks categorically. On the contrary, these prosodic features, opposed to segmental features like nasal or rounded, do not function in a binary fashion. It seems at this stage of research that hypothesis (1) obviously was not justified.

When the tonal manifestation of the focus accent is concerned, we know that the size of the rise may vary considerably in speech. However, a survey of the use of focus accent, its distribution and the variations in manifestation, in spontaneous speech of different varieties of Swedish is badly needed. This applies also to chunking. We know that silent interval, low Fo and final lengthening, single or combined, are strong cues to phrase boundaries. Unfortunately, we still know very little about the rôle that voice quality, intensity and perhaps other features play for chunking.

References

[1] Bannert, R. (1993), Macro-prosodic organisation of spoken Swedish: Some preliminary observations on focus accent and chunking. *Reports from the Department of Phonetics, Umeå University, PHONUM* 2, pp. 21–39.

[2] Bannert, R. (1994), Listeners' Identification of Prominence and Chunking in Spoken Swedish: Variation and Consistency. *Working Papers, Department of Linguistics, Lund University* 43, pp. 10–13.

[3] Bannert, R. (1987), From Prominent Syllables to a Skeleton of Meaning: a Model of Prosodically Guided Speech Recognition. *Proceedings of the 11th International Congress of Phonetic Sciences, Tallinn, USSR.* 2, pp. 73–76.

[4] Bannert, R. (1991), Automatic Recognition of Focus Accent in German. *Journal of Semantics* 6/7. (Special Issue: Views of Focus. Proceedings of a Workshop on Focus, Intonation, and Semantics).

[5] Bannert, R. (1995), Variations in the perceptual modelling of macro-prosodic organization of spoken Swedish: prominence and chunking. *Reports fromthe Department of Phonetics, Umeå University PHONUM* 3, pp. 7-30.

[6] Bruce, G. (1977), Swedish Word Accents in Sentence Perspective. *Travaux de l'Institut de Linguistique de Lund XII.* Lund: Gleerup.

[7] Gårding, E. (1967), Prosodiska drag i spontant och uppläst tal. In: Holm, G. (ed) *Svenskt talspråk*, pp. 40-85. Stockholm: Almqvist och Wiksell.

PITCH AND NON-PITCH CUES TO WORD STRESS IN CZECH

Christine Bartels
University of Massachusetts, Amherst, U.S.A.

ABSTRACT

This study investigated the role of pitch, duration, amplitude, and vowel formants in word stress in Czech. It was found that stress and pitch peak tend to diverge, in that the latter occurs more often on the syllable following the stressed, first syllable than on the stressed syllable itself. Further, the placement of the pitch peak is dependent on vowel quantity and vowel height: both long and high vowels are relatively favored. Among the remaining cues, syllable duration outranked amplitude and vowel formant ratios.

INTRODUCTION

Czech is traditionally described as a stress-accent language having fixed word stress on the first syllable, independent of vowel quantity [1,2]. It shows high and low pitch accents and characteristic phrasal intonation contours but does not use tone for lexical contrast. Pitch accents generally fall on the initial, stressed syllable. This correlation is not absolute, however; the pitch accent is frequently placed later in the word, at times leading to the perception among non-native listeners that this later syllable is the more prominent one [1]. The production study reported here investigated the role of two contextual factors, vowel height and vowel quantity, in conditioning such divergent placement of stress and pitch peak in Czech. These factors were chosen because the informal observation is that high and long vowels seem particularly effective in attracting the pitch peak. The study also examined whether other acoustic measures serve as reliable correlates of word stress in Czech. In the non-tonal stress languages English [3] and German [4], duration and amplitude may have equal or greater importance than F_0 in marking the stressed syllable, and spectral characteristics also contribute to conveying word stress in English [5]. Even within a single language, different levels of the prosodic hierarchy may put different weight on these acoustic cues in coding prominence [6,7]. By looking at a stress language that is untypically unreliable in its use of pitch cues, this study aimed to contribute to the growing experimental evidence for the phonetic diversity hiding behind the traditional notion of "stress."

METHOD

Nine native speakers of Czech, five women and four men, were asked to produce contextually embedded trisyllabic nonsense words of the form /nV(:).nV(:).na/. The vowels in the first and second syllable were either [i] or [a], and either short or long. Each speaker read these words twice in different random orders, with each word embedded in the frame sentence *Řekni ____ ještě jednou* ('Say ____ once more'), which places main sentence prominence on the test word. The recorded test words (288 tokens in all) were digitized at 10 kHz and the following acoustic measures taken from waveform, F_0 contours, or spectrograms of the first two syllables: syllable and vowel duration, average, peak, and total rms amplitudes, and average *F*1 and *F*2 frequencies from the central 50 ms of the vowel's steady state. Vowel quantity (short *vs.* long), vowel quality ([a] *vs.* [i]), and where appropriate, syllable position served as the independent variables in separate repeated measures ANOVAs for each of the acoustic measures listed above, as well as for certain first/second syllable ratios. To ascertain that the absolute timing of the F_0 peak relative to word onset did vary and that its variable alignment was not just a reflection of variation in syllable length, a simple regression analysis was carried out with duration of the first syllable as the independent variable.

RESULTS

Pitch excursions were rather small throughout for these Czech speakers. Only five out of nine subjects, referred to as the H* group below, showed clear peaks in the test word; the others did not exhibit a consistent pattern. Even within the H* group, speakers varied considerably in the range of timing produced.

If a token showed a distinct pitch peak, it occurred no later than the second syllable. Within the first two syllables, the absolute timing of the F_0 maximum with respect to word onset and syllable boundaries varied widely, but peaks rarely occurred before the second half of the first vowel and were always preceded by a distinct, gradual rise.

Shift of the F_0 maximum to the second syllable was the rule rather than the exception in these data: In 58.3% of the tokens, the ratio of the F_0 maxima for first and second syllable was lower than predicted based on average peak values for the vowels and quantities involved in a given token; thus, only 41.7% of the time did the overall pitch peak correctly identify the stressed syllable. Averaging F_0 maxima across subjects for first and second syllable, no significant difference was found. Averaging across the H* group only, the mean values were 193.7 Hz and 197.4 Hz, respectively, a minor difference, but one that favors the second syllable.

When tested across all nine speakers, in the first syllable but not the second peak height correlated significantly with quantity [$F(3,51) = 8.736$; $p = .0001$] and vowel quality [$F(3,51) = 4.574$; $p = .0065$], being higher for long vowels and for [i]. For long but not short first syllables, peak height of either syllable correlated with the vowel height of the first vowel [$F(9,153) = 2.672$; $p = .0066$). The same test carried out for the five H* speakers alone yielded an additional main effect for syllable position itself [$F(1,9) = 9.345$; $p = .0136$], with the F_0 maximum being slightly higher for the second syllable (see above), but the dependence of the second-syllable maximum on the height of a long first vowel was no longer significant.

The first/second syllable ratio of F_0 maxima also depended strongly on both vowel quantity and vowel quality. On the one hand, a long first vowel but not a short one attracted the overall pitch maximum [$F(3,51) = 7.365$; $p = .0003$]. On the other, a high second vowel following a low first vowel, i.e. the combination /na(:)ni(:)na/, shifted the peak to the second syllable [$F(3,51) = 3.797$; $p = .0237$]. In cases of conflict, neither effect dominated consistently. When the same analysis was carried out for the H* group only, similar main effects were found: a long first vowel attracted the pitch more than a short one [$F(3,27) = 17.593$; $p = .0001$], and a high first vowel did so more than a low one [$F(3,27) = 7.094$; $p = .0012$]. However, for this group the ratio was >1, i.e., the overall pitch maximum was found to fall on the first syllable, only if that syllable contained a long [i].

The effects of vowel quantity and vowel quality on F_0 syllable averages were similar, but because of the word-initial rise from a lowpoint in H* peak contours, this measure yielded an even

lower score as a stress cue: only in 21.5% of tokens did the ratio of F_0 averages, adjusted for vowel properties, favor the first syllable.

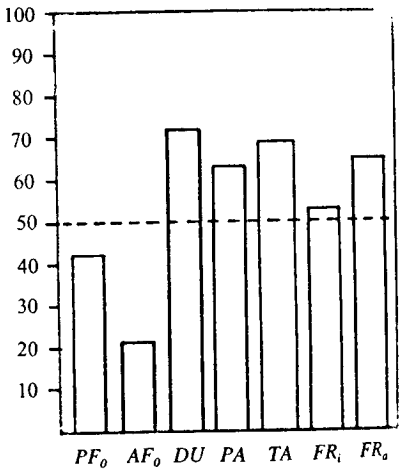
Among the other stress cues examined, syllable duration was the one most frequently employed. (Vowel duration was a much less reliable measure.) The average syllable durations, by vowel type, were: short [a], 187.4 ms; long [a], 330.1 ms; short [i], 174.7 ms; long [i], 278.1 ms. The ratio for first *vs.* second syllable for a given token based on these values was exceeded in 71.5% of all tokens.

Average syllable amplitude and peak syllable amplitude also played a role in cuing for stress. Adjusted for vowel height and length effects, they favored the first syllable in 59.0% and 63.9% of all tokens. Because of the bias in syllable durations, total syllable amplitude, at 68.1%, did even better.

Formant values contributed negligibly if the vowel involved was [i]: the ratio *F2/F1*, expected to increase for more "hyperarticulated" forms of the vowel and in fact doing so in longer vowels here, exceeded average values for these speakers only 52.8% of the time in first-syllable [i]'s. Formant values contributed somewhat more in [a]'s: both *F1* and *F2* were significantly higher in first than in second syllables, while at the same time, in a stronger, somewhat opposing trend toward "hyperarticulation," the second formant decreased with vowel length. Thus both *F1* alone and the mean of the two formants exceeded average values 59.7% of the time in first-syllable [a]'s, *F2* alone exceeded the average 61.1% of the time, and the ratio *F2/F1* stayed below the average in 63.9% of all tokens containing first-syllable [a]'s.

Figure 1 summarizes these results.

Figure 1. Percentage of tokens in which the acoustic cues measured correctly identified the stressed syllable. PF_0, peak F_0; AF_0, average F_0; DU, syllable duration; PA, peak amplitude; TA, total amplitude; FR_i, formant ratio [i]; FR_a, formant ratio [a].



DISCUSSION

This study confirms the listener's impression that pitch, in terms of F_0 maximum and average, is not a reliable cue to stress in Czech; in fact, in the type of data examined here, it shows a tendency to peak *after* the stressed first syllable. As suspected, this tendency turns out to be strongly dependent on segmental factors such as vowel height and vowel length, in that both long and high vowels attract pitch.

Of the typical non-pitch cues to stress, i.e. duration, amplitude, and spectral characteristics of vowels, only duration comes close in Czech to the role it plays in English, where it appears to be the least reliable cue [3]. Thus the results of this study are in agreement with the general impression that Czech has weak stress accents only, compared to other Slavic languages. However, duration and amplitude outrank pitch; this fact matches the conclusions of an early perception study conducted with Czech listeners [8].

It appears plausible, though, that in the H* speakers at least, the prolonged word-initial *rise* in F_0 has become relevant; this cue is perceptually salient but is not usually considered in quantitative studies. In a range of languages, for instance Swedish [9] and Japanese [10], dialects differ as to the temporal alignment of pitch accents and pitch movements with syllable structure, usually in the direction of relative delay of movement endpoints. Thus one might speculate that to the extent that Czech speakers use a distinctive pitch accent at the word level, once the strong force of segmental effects has been accounted for, the temporal placement of this peak presents a compromise between the need to align it with the word stress and the attempt to maximize the salience of the preceding rise.

ACKNOWLEDGMENTS

I gratefully acknowledge the technical assistance and helpful comments I received from John Kingston, as well as the discussions I had with Lisa Selkirk and John McCarthy. Thanks to members of the Institute of Formal and Applied Linguistics and the Institute of Theoretical and Computational Linguistics of Charles University, Prague, as well as to the members of the Institute of Experimental Biology of the Czech Academy of Science, Olomouc, for their willingness to serve as subjects, and to Julia Hirschberg, Lou Conover, and Yuzo Fujikura for their active support. This work was sponsored in part by the U.S.-Czech/Slovak Science and Technology Joint Fund, project no. 92058.

REFERENCES

[1] Broch, O. (1911), *Slavische Phonetik*, Heidelberg: Carl Winter's Universitätsbuchhandlung.

[2] Jakobson, R. (1962 [1926]), "Contributions to the study of Czech accent," in *Selected Writings I: Phonological Studies*, The Hague: Mouton, pp. 614-625.

[3] Beckman, M. (1986), *Stress and Non-Stress Accent*, Dordrecht: Foris.

[4] Kohler, K. (1991), "Terminal intonation patterns in single-accent utterances of German: Phonetics, phonology, and semantics," in K. Kohler (ed.), *Studies in German Intonation*, Arbeitsbericht Nr. 25, Institut für Phonetik, Kiel University.

[5] de Jong, K. (1995), "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *J. Acoust. Soc. Am.* 97, 491-504.

[6] Beckman, M., and J. Edwards (1994), "Articulatory evidence for differentiating stress categories," in P. Keating (ed.) *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology 3*, Cambridge: Cambridge University Press, pp. 7-33.

[7] Shattuck-Hufnagel, S., M. Ostendorf, and K. Ross (1994), "Stress shift and early pitch accent placement in lexical items in American English," *J. Phonetics* 22, 357-388.

[8] Janota, P. (1967), "Perception of stress by Czech listeners," in *Proceedings of the Sixth International Congress of Phonetic Sciences*, pp. 457-461.

[9] Bruce, G. (1983), "Accentuation and timing in Swedish", *Folia Linguistica* 17, 221-238.

[10] Pierrehumbert, J., and M. Beckman (1988), *Japanese Tone Structure*, Cambridge, Mass.: MIT Press.

LEXICAL FILTERING BY MEANS OF PROSODIC INFORMATION

F. Béchet (*), P. Langlais (**), H. Méloni (*)
(*) Laboratoire d'Informatique d'Avignon (LIUAPV)
33, rue Louis Pasteur - 84000 Avignon - France
(**) IDIAP
Case Postale 592 CH 1920 - Martigny - Suisse

ABSTRACT

We present in this article a study on the integration of prosodic information in a lexical access module. Our approach consists, in a first step, to verify the pertinence of microprosodic features contained in fundamental frequency, duration and intensity parameters. We have realised, in order to do that, an inventory of relevant features which have been the subject of many studies by the past. Then, we describe their performance (in filtering and/or verification) when they are measured in an automatic way. As a conclusion, we will present an implementation of an efficient filter using the fundamental frequency parameter.

INTRODUCTION

Many studies point out the essential role of prosody in the linguistic code and mainly in emotional, pragmatic, semantic and syntactic areas. That does not mean however that prosody is not involved into lower levels. Di Cristo [1] has shown the importance of microprosodic information in French already underlined for other languages by earlier studies. Even if our language has no functional stress pattern at word level as it is the case in English, studying the integration of some prosodic features - such as duration, intensity and fundamental frequency variation - in a task of lexical filtering appears fully justified to us.

This study takes place in the general framework of word recognition processes developed in LIUAPV. So as to determine the pertinence and robustness of the prosodic features studied, we plan to add a prosodic lexical filter to the lexical access module of the SPEX system [2]. We deal here with problems posed by the recognition of isolated words from very large vocabularies. All modules developed in this project have a "knowledge based" approach. One of the main purposes is to propose an alternative to systems using statistical methods with a large training corpus. In our system the training needed for every new speaker is reduced to the realisation of 30 words chosen for containing all the French phonemes in various contexts introducing little distortion.

As we already have a lexical access module made of filters progressively reducing the dictionary of possible words, the questions we want to answer, at the beginning of this study, are :

- Is taking into account additional prosodic information likely to improve the process described ?
- In the affirmative, what is accurately the relevant information ?
- Is the latter robust enough to allow its efficient integration into our lexical module ?

We will first briefly present the lexical access module and the prosodic features studied.

THE SPEX SYSTEM

The SPEX system operates on two levels:

- a set of phonetic modules including an Acoustic Phonetic Decoding process (APD) [3] producing from speech signal and speaker references a lattice of valuated phonetic hypotheses
- a lexical access module which filters the global lexicon to propose a subset of candidates to the evaluation process.

The goal of a lexical access module is to find a correspondence function which links the phonetic units recognised with the lexicon. The first operation is choosing the kind of unit suitable to make the link with the lexical items. The phenomena of insertion, deletion and substitution which appear in the phonetic lattice lead us to think that the phoneme unit is not a realistic choice because of the insufficient performance of the bottom-up Acoustic Phonetic Decoding system. In fact there are too many uncertainties to allow a direct access to some parts of the lexicon.

Therefore the choice has been made to use macro-sets representing sounds which have distinctive acoustic features. The advantage of such a representation is to put a structure on the information contained in the phonetic lattice. The number of possible paths inside the lattice for identifying a word is then reduced.

Our lexical access method consists in representing the lexical data and the phonetic hypotheses in a common structure. This structure will allow us to select, in a bottom-up way, some lexicon items. By making this structure more accurate till the phonetic description of each item is complete, we progressively reduce the number of lexical hypotheses in order to give a cohort of valuated items as a probable solution.

The last step in the recognition process consists in sharply evaluating the hypotheses left. To this purpose, by means of spectral distances and contextual articulatory features, we confront the calculated phonetic decomposition of each item of the cohort remaining with the effective realisation of the sounds by the speaker.

PROSODIC FEATURES STUDIED

All the prosodic features chosen have been studied on large test corpora (from 500 to 1000 words pronounced by several speakers).

Duration

Thanks to numerous studies concerning the temporal aspect of speech, we know that the acoustic symptoms of this parameter are governed by multiple factors. Consequently, we are immediately confronted with two difficulties when we want to use them in an automatic process :

- How can we segment a speech signal into discrete units (in our case phonemes) ?
- What precision can we expect from our measures?

We have to study the variations of vowel duration by taking the duration measures produced by the lexical module. A precise study of this parameter can be found in [4]. Here is a summary of this study.

Among all the factors which influence intrinsic vowel duration, it seems that only a few can be observed - at least by our techniques - beyond average values. One can however retain that :

- an oral/nasal distinction can be made, at least partially, with a low error rate,
- the position of the vowel strongly influences its length, but the phenomenon is in no way easily localizable at the end of the word,
- the influence of the consonant to the right is not easily measured,

As a conclusion, it seems that intrinsic vowel duration is not reliable enough to be used in our system (except oral/nasal vowel distinction) because of its lack of robustness - largely due to its bad automatic extraction. The total output of these features in a top-bottom utilisation is rather weak. Therefore it does not seem judicious to integrate them for the moment.

Variation of fundamental frequency

It is often argued that fundamental frequency can be used with great benefit in a segmentation process. In spite of this opinion, it is rarely used in recognition systems for several reasons, the principal one being the lack of reliability of *f0* measures.

All intrinsic and co-intrinsic variation factors of vowel fundamental frequency can be reduced to the articulatory model of the vowel and to the voice/voiceless characteristic of the previous consonant.

About consonants, the shape of the *f0* curve can provide indications for the distinction between obstruent and non obstruent consonants. We have integrated in our lexical access module a filter based on a Bayesian decision of the obstruent/non obstruent nature of inter-vocalic consonants. This decision, calculated from the distributions of fundamental frequency measured on our corpora, allow us to filter about 15% of our word cohorts with an average gain of two position (when the word pronounced is not classified first). However the rejection rate is about 7%.

Intensity

Intensity is without doubt the least studied parameter of prosodic research, although it is by far the easiest to extract from the speech signal. The few studies on this parameter [5], however, point out the following result : the global intensity of a vowel generally seems weaker when it is preceded by a voiceless consonants ; low vowels generally have a higher specific intensity than high vowels (with a minimum for vowel /i/ and a maximum for vowel /a/ and /ɔ/).

We have thus developed a filter based on a decision made from the distribution of initial vowels /i/ and /a/ in our corpora. The error rate is low, but the filtering rate is not very efficient.

Voice/voiceless discrimination

The fundamental frequency parameter allows us to distinguish voiced consonants from others. We know, however, that a voice/voiceless distinction from the speech signal is difficult to obtain in all conditions. We measure this parameter with an algorithm based on the Amdf method, which gives good results. The voice/voiceless decision is taken according to the shape of the Amdf curve calculated on every signal frame. The results obtained on our test corpora allow us to consider efficient the use of this parameter for our lexical filtering task.

The lack of robustness of most of the microprosodic features examined leads us to implement, at first, a lexical filter using the voice/voiceless decision curve.

IMPLEMENTATION

We break up the problem in two stages :

• A first filter works before the recognition process in a bottom-up way. To this end it uses the phonetic chain associated to each word of the lexicon. This chain represents the "usual" pronunciation of these words. We have now to eliminate the candidates whose "theoretical" voice pattern does not match with those measured on the speech signal.

• The second step consists in filtering in a top-bottom way the resulting cohort produced by the lexical process by eliminating the words whose calculated voice pattern is not compatible with those of the speech signal. At this step we have a number of hypotheses about the phonetic chain and its temporal position.

Filter 1

The voice pattern of the signal is obtained from the voice curve calculated with the fundamental frequency variation curve. A "theoretical" voice pattern for every item of the lexicon is made according to the following technique :

• Very few words include a sequence of three consonants. We do not take into account their possible consonantal assimilations.

• When we have a sequence of two consonants, the voiceless consonantal assimilations are ignored. The temporal alignment is not yet known, so these phenomena cannot affect the voice pattern of the word.

• We then consider the possible voice consonantal assimilations which can affect the voice pattern.

• Finally we take into account the fading of the /ə/ at the end of a word.

We have tested the efficiency of our filter on the corpus AviLex (700 words pronounced by six speakers) [2] previously used for testing the lexical access module, with the same lexicons of 15 000 and 20 000 words. The results of table 1 show a filtering rate of 60% for an error rate of 3%. This filtering rate is good, compared to the simplicity of the techniques used.

Table 1 : result of filter 1

speaker	filtering	errors
fb	60%	2.9%
hm	60%	2.1%
ts	59%	3.4%
pg	59%	1.2%
lc	59%	3.7%
si	59%	2.7%
all	59.3%	2.6%

We wish to specify that these results are obtained with a speaker independent algorithm. An error rate under 1% - for the same filtering rate - can be reached when we determine a voice threshold specific for each speaker. Although we think that it is possible to automatically determine this threshold during the training phase of our system, we did not proceed further in this direction because of the efficiency of the global algorithm.

Filter 2

This filter is used at the end of the recognition process whose output is a valuated cohort of 50 to 150 words containing the pronounced item.

Unlike the first filter, it operates with the phonetic chain supposed and its temporal alignment. Thanks to this information, it is then possible to sharply filter the word cohort by integrating all the consonantal assimilation and /ə/ elision rules.

Table 2 presents the results obtained for all the speakers of the AviLex corpus. We obtain an average filtering rate of 20% on the word cohort with an error rate of about 3%. The errors can be explained by a bad voice decision (30% of the errors) or by a mistake in the phonetic alignment proposed by the lexical level of the system. The average gain - for all the speakers - is about three positions up when the pronounced word is not first in the cohort.

Table 2 : result of filter 2

speaker	filtering	errors	gain
pg	20.9%	2.1%	3.5
fb	17.5%	2.2%	2.7
ts	21%	2.7%	3.7
si	16.9%	2.2%	3
lc	17.8%	1.5%	2
all	18.8%	2.1%	3

CONCLUSION

The study of the integration of microprosodic features in a lexical access module suggests the following comments.

Most of the various filters realised had insufficient robustness. Without denying the important role of the microprosodic features in the discrimination of sounds, it seems that the automatic extraction of these features leads to a loss of precision in their measurements. Because of this situation, we use average values for all the features and the filtering rate obtained is rather disappointing.

Nevertheless the results obtained by our filter using the voice/voiceless detection curve justify the use of robust features in a lexical access process (in a top-bottom or bottom-up way). Even if it is pretentious to talk here of a prosodic treatment, this method has the advantage of presenting good results which can be easily integrated into an automatic speech recognition system.

REFERENCES

[1] A. Di Cristo ; *De la microprosodie à l'intonosyntaxe* ; Université de Provence, 1985.

[2] F. Béchet, *Système de traitement de connaissances phonétiques et lexicales : application à la reconnaissance de mots isolés sur de grands vocabulaires et à la recherche de mots cibles dans un discours continu*, Thèse de l'Université d'Avignon et des Pays de Vaucluse, 1994.

[3] P. Gilles, *Représentation et traitement de connaissances acoustiques et phonétiques pour la reconnaissance de la parole*, Thèse de l'Université d'Avignon et des Pays de Vaucluse, 1993.

[4] P. Langlais, *Promenade légère dans les allées prosodiques du jardin de la parole* ; Thèse de l'Université d'Avignon et des Pays de Vaucluse, 1995.

[5] M. Rossi, *Interaction des glissements d'intensité et des glissements de fréquences*; XIVth International Conference on Acoustics, 1976.

THE USE OF PROSODIC INFORMATION IN WORD RECOGNITION IN MODERN STANDARD ARABIC

Sami Boudelaa
Laboratoire de Phonétique, Université PARIS 7
FRANCE

ABSTRACT

A semantic priming experiment investigated the effect of lexical stress during auditory word recognition in Arabic. In minimal stress pairs, lexical decision was inhibited only through rightward stress movements. In common words, stress shifts were adverse in both directions. The results are explained in terms of stress pattern frequency and syllable weight.

INTRODUCTION

Stress, the relative prominence of one syllable within a word [1, 2] is said to be lexical when it is functionally distinctive. Attempts to detail its influence present a rather confusing picture.

It seems that in English prior information as to the number of syllables and lexical stress pattern of a target word does not improve lexical decision performances. Also, mis-stressing inhibits word recognition only if a canonically strong-weak (/SW/) stress pattern is realized in a /WS/ version [3]. More important still, minimal stress pairs such as "forbear/forbear" behave like homophones, suggesting that lexical stress information is not used to constrain lexical access [4]. Likewise, the mis-stressing of pairs like "contract-contract" does not impede word processing, eventhough it involves a vowel quality change [5].

However, positive evidence regarding the influence of lexical stress on word recognition also exists. For instance, English listeners' identification of an ambiguous initial segment is biased in the midrange of a speech voicing continuum by stress information [6]. Also, the detection of mispronounced targets is greater in stressed than in unstressed syllables[7], and mis-stressing results in slower shadowing responses, whether a vowel quality change is involved [8] or not [9]. Finally, gating evidence shows that the words suggested on the basis of gated information differ depending on whether the word is /SW/ or /WS/ [10].

Given the inconclusive results from earlier studies, it would be interesting to provide additional cross-language information from semantic priming regarding the potential effects of lexical stress on spoken word recognition.

In Modern Standard Arabic (MSA) stress pattern can have a lexically distinctive function in the sense that there are few minimal stress pairs [11, 12, 13]. Such pairs consist exclusively of three-syllable words and are either /SW/ or /WS/, final syllables being almost always extrametrical in this language unless superheavy [14]. For instance, the sequence /wæsʕafa/, with a /SW/ stress pattern means "he described", but with a /WS/ stress pattern, /wæsʕafa/ means "it cleared up". Being semantically different, members of such pairs are supposed to be related to different words on the representational level [4]. The /SW/ version is related to the word /ʃæræħæ/ (i.e., he explained), while the /WS/ one is related to /ræ:qæ/ (i.e., it became brighter). A contribution of lexical stress to the process of word recognition in MSA can be demonstrated, if a member of a minimal stress pair is found to facilitate only the recognition of the target related to it. On the other hand, if stress plays no role, then minimal stress pairs should behave like homophones [4, 5]. In order to further define the role of lexical stress in word recognition, it would be of interest to examine the perceptual effects of mis-stressing /SW/ and /WS/ MSA common words, that is words which are not members of a minimal stress pair (e.g., /kætæbæ/-/ʃæ:hædnæ:/ (i.e., "he wrote, they saw" respectively). Should mis-stressing have an adverse effect on lexical access, then the correctly stressed versions of common words should facilitate related targets, while the incorrectly stressed versions should not. It is worth noting

that stress manipulation in MSA has no consequence at the segmental level [15], thus allowing a better assessment of lexical stress effects than a language like English in which stress shifts usuallly alter vowel quality [4].

A PRIMING EXPERIMENT

The role of lexical stress during auditory word recognition in MSA has been tested in a semantic priming experiment in which subjects made a lexical decision for a target which was or was not related to a preceding prime word. Preliminary control studies were run to construct reliable material relative to the associative relations between primes and targets, and to determine an unprimed baseline lexical decision time.

METHOD

Subjects

Twenty four student volunteers aged between 23 and 34 took part in the experiment. They all were native Arabic speakers with no known history of hearing loss or speech disorder.

Materials

The materials consisted of two sets of three-syllable words controlled for frequency [16]. The first set comprised 18 quadruplets of which the first item was a /SW/ or /WS/ minimal stress pair member. Each member of minimal stress pairs served as a prime either to a target semantically related to it (R1), or a target related to the second member of the pair (R2) or to a control word (C), which was matched to the prime as closely as possible on syllable length, frequency of occurrence, word class and polysemy. The second set consisted of 18 triplets of which the first item was a /SW/ or a /WS/ common word token realized in a correctly stressed (CS) or a mis-stressed version (MS). Mis-stressing resulted when stress was shifted either to the right in the case of a /SW/ word, or to the left in the case of a /WS/ one. The CS and the MS versions of such words were used to prime semantically related and control targets (C). In addition, 126 words were selected to serve as primes to nonword targets formed by changing one to two phonemes across all possible positions in the original 126 words. Four lexical decision lists were prepared each containing 62 to 64 items half of which were non-word targets. The other half consisted in word targets primed either by a member of a minimal stress pair, a correctly stressed common word or a mis-stressed common word. Stimuli were recorded in a sound-treated room using a Sony double-deck cassette (TW320) and a microphone Vivanco (EM 238) to be digitized later at a sampling rate of 10 kHz and a 12 bit resolution.

Procedure

Subjects, tested individually in a quiet room heard, the stimuli over a pair of headphones. A practice set comprised 24 trials half of which were non-words. The interstimulus interval was 100 ms, while the inter-trial interval was 1s. Stimuli were presented in two blocks containing two experimental lists each. A five-minute pause separated the presentation of the two lists within a block which was presented to half of the subjects. The prime-target pairs were counterbalanced across the lists and their presentation was randomized for each block. The same test word never appeared twice in the same list. Subjects had to respond "word" or "nonword" as quickly and as accurately as possible by pressing one of the two appropriately labelled response keys which were counterbalanced across subjects. The presentation of stimuli and collection of data were controlled on-line by a Toshiba T 5200, using a da_tr program (Hallé 1991). Response times were measured from the acoustic offset of the target word.

RESULTS

Minimal stress pair analysis

Subjects' reponses included a low error rate -3%- both for minimal stress pairs and common words, so the analyses to be presented below concern RT's only. /WS/ words were longer in duration and yielded longer RTs than /SW/ words. fig.1. displays subjects' mean RTs. A two-way ANOVA -by subject F1 and by items F2- showed that the main effect of Stress was not significant [F1(1,23) = 0,49, p = ,5., F2(1,48) = 0,69,p =,5], reflecting the absence of difference in the processing of targets presented after /SW/ and /WS/ primes.There was, however a significant

main effect of Relation [F1(2,23) = 28, p = ,05., F2(2,48) = 96, p = ,05]. The interaction between the two factors was also significant in both analyses [F1(2,48) = 21, p < ,05., F2(2,48) = 18, p < ,05], with R2 responded to as quickly as R1 when the prime was /SW/. When the prime was a /WS/ item however, R1 was responded to significantly more quickly than R2 whose response time did not excel that of the control word C.

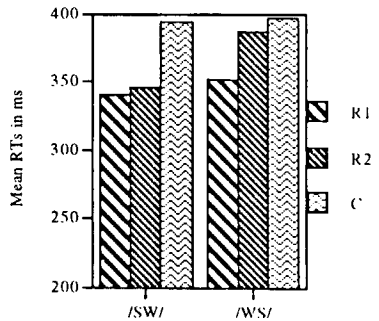


Fig.1. Mean lexical decision times in ms. R1= prime and target are semantically related, R2 = the target is primed by the member of the stress partner to which it is not related, C = control word.

In other words, while the target /ræ:qæ/, which is related to the /WS/ /wæsʕafa/ can be facilitated both by the /WS/ and the /SW/ versions of the sequence /wæsʕafa/, the target /ʃæræhæ/ which is semantically related to the /SW/ member of the minimal stress pair was facilitated only when preceded by the relevant priming stress partner.

Accordingly, our data do not concur entirely with those of Cutler [4], who argues that there is little premium in computing lexical stress on-line on the basis of her finding that English minimal stress pairs behave like homophones. Indeed, it would be counterintuive to sustain such an idea in MSA for the following reason: Lexical stress conveys morphological information in the sense that a stressed syllable always contains at least one segment belonging to the root morpheme, and root morphemes have a special status in MSA as they are associated with a semantic load the knowledge of which is crucial to the understanding of all the morphologically complex words. So, the failure to observe any leftward stress movement effects in minimal stress pairs may be due to the fact that the movement is between two syllables of equal weight. Furthermore, the /SW/ stress pattern is of higher frequecy because in MSA lexical stress assignment proceeds from right to left and the syllable on the right is more often than not an unstressable syllable [17].

Common Word Analysis

Mean lexical decision times in ms are displayed in Fig. 2. A two-way ANOVA revealed significant main effects of Stress [F1(1,23) = 133, p< ,05., F2(1,48) = 2,3, p<,05] and Relation [F1(2,23) = 7,5, p= ,05., F2(2,48) = 12, p= ,05]. The interaction was not significant, however [F1(1,23) = 0,49, p = ,5., F2(1,48) = 0,8, p = ,5].

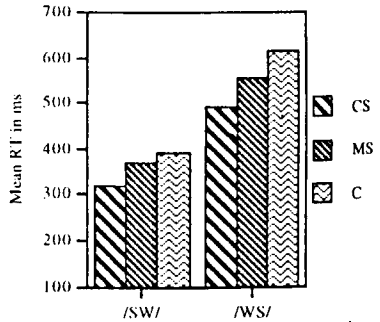


Fig. 2. Mean lexical decision times in ms. CS = a correctly stressed prime followed by a related target. MS = a mis-stressed prime followed by a related target. C = a control word.

The common word data show that lexical decision is seriously impeded both when stress is moved leftwards and rightwards. A /SW/ common word like /kætæbæ/ fails to prime a related target when it is realized in an unorthodox /WS/ stress pattern /kætæbæ/. Similarly, a canonically /WS/ common word like /ʃæ:hadna/ is of little facilitatory affect when mis-stressed as /ʃæ:hadna/.

This result shows that stress movements in both directions have an adverse effect on word processing. This may be explained as follows: Mis-stressing a /SW/ common word amounts to replacing a frequent stress pattern by a less frequent one, while mis-strsessing a /WS/ common word involves a stress shift from a heavy syllable to a light syllable, that is from a CVC to a CV. So in both cases word processing is impeded. We are tempted to say that the effects of syllable weight and stress pattern frequency are additive, although our data do not address this question directly.

CONCLUSION

Two key outcomes emerge from the experiment: First, priming is unaffected by *leftward stress movements* in minimal stress pairs. Second, both *leftward and rightward stress movements affect priming in common words*. It is suggested that when stress movements involve a shift between syllables of equal weight and when it results in a more frequent stress pattern, it is without effect. But when it is from a heavy syllable to a light one, or when it substitutes a less frequent stress pattern for a dominant one, a significantly less priming effect results.

Overall lexical stress is important in MSA as it conveys morphological information that is crucial to the meaning of the word [18]. Moreover, the *reduced variability of syllable structure*, the *ease with which syllable boundaries can be located and the interaction between syllable structure and lexical stress* all make the drawing on lexical stress in MSA a real premium.

REFERENCES

[1] Bolinger, D. (1986), *Intonation and its Parts*, Stanford University Press.

[2] Jassem, W. & Gibbon, D. (1980), Redfining English accent and stress. *J. Int Phon Ass*, 1-2, 2-16.

[3] Cutler, A. & Clifton, C. (1984). The use of prosodic information in word recognition. in H. Bouma & D. Bouwhis. (eds).*Attention and Perfomance: X. Control of Language Processes*. 183-196 Hillsdale, NJ.

[4] Cutler, A. (1986), Forbear is a homophone: Lexical prosody does not constrain lexical access, *Lg. Sp.* 3: 201-219.

[5] Small, L., Simon, S. & Goldberg, J. (1988), Lexical stress and lexical access: Homographs versus nonhomographs. *Per. Psychophysics*, 3, 272-280.

[6] Connine, S., Clifton, C. & Cutler, A. (1987), Effects of lexical stress on phonetic categorization, *Phonetica*, 44, 133-146

[7] Cole, R., & Jakimik, J. (1980). How are syllables used to recognize words?, *J.Acoust Soc Am*, 67, 965-970.

[8] Small, L. & Bond, Z. (1982), *Effects of mispronunciations on lexical access in continuous speech perception*. Paper presented at the meeting of The American Speech-Language Hearing Association, Toronto.

[9] Slowiaczek, L. (1990), Effects of lexical stress in auditory word recognition, *Lg.Sp.* 33, 47-68.

[10] Van Heuven, V. (1988), Effects of stress and accent on the human recognition of word fragments in spoken context: Gating and Shadowing, *FASE Symposium*, 811-818, Edinburgh,

[11] Abdou, D. (1969). *On Stress and Arabic Phonology: A Generative Approach*. Beyrouth.

[12] Brame, M. (1971), Stress in Arabic and generative phonology, *Foundations of Lg.* 7, 556-591.

[13] Rajouani, A., Chiadmi, D., Najim, M. & Ouadou, M. (1988), L'accent lexical en arabe: *PACS*, 43-72.

[14] Badri, K. (1982), *La Linguistique Appliquée*, Saoudi Arabia.

[15] Belkaid, Y. (1984), Les voyelles de l'arabe standard moderne: analyse spectrographique, *TIPS* 16, 217-240.

[16] Kouloughli, D. E.(1991), *Basic Lexicon of Modern Standard Arabic*, L'Harmattan, Paris.

[17] Angoujard, J. (1990), *Metrical Structure of Arabic*, Foris Dordrecht.

[18] McCarthy, J. (1979), *Formal Problems in Semitic Phonology and Morphology*, Unpublished PhD, MIT.

THE COMBINED EFFECTS OF PROSODIC VARIATION ON JAPANESE MORA TIMING

Ann R. Bradlow, Robert F. Port, and Keiichi Tajima
Indiana University, Bloomington, Indiana, U.S.A.

ABSTRACT

Previous research has shown that the mora functions as a consistent timing unit in Japanese, such that there is a linear relationship between total word duration and the number of moras in the word. The present study extends the examination of Japanese as a mora-timed language by investigating the combined effects of variation in overall speaking rate, sentence-level focus, as well as number of moras on total word duration. The data indicate that Japanese timing is consistently constrained by the mora-timing principle. These findings have implications regarding the function of this articulatory-acoustic regularity from the listener's point of view.

INTRODUCTION

In both traditional and contemporary work, Japanese is described as a mora-timed language. As evidence of mora-timing in Japanese, traditional accounts point to the mora-based metrical structure of traditional Japanese poetry, to the *kana* orthographic system in which each symbol represents a mora-sized unit, and to the general native-speaker awareness of the mora ("onsetsu" or "haku") as a unit of equal timing. More recent instrumental analyses have focused on finding acoustic evidence of Japanese mora timing. Although Beckman [1] found that moras of varying segmental content did not exhibit the expected durational consistency, Port et al. [2] and Han [3] showed a linear relationship between the number of moras in a word and the total word duration. Data from these latter studies indicated that the mora-based regularity of Japanese timing is expressed at the level of word-sized units: words of a given number of moras all achieve a constant duration which is directly proportional to the number of moras in the word. Both of these studies revealed durational compensation at the "sub-moraic" level in order to achieve a target total word duration that is consistent with this mora-timing principle. The present study was designed to explore the effect on mora-timing of prosodic variation at the "supra-moraic" level. Specifically, we investigated the combined effects of variation in overall speech rate and sentence-level focus on the duration of words with varying numbers of moras. Our expectation was that the linear relationship between total word duration and number of moras in the word would persist even under conditions that might otherwise be expected to perturb this regularity in the time dimension. Such a result would provide further evidence regarding the extent to which Japanese mora timing is regular and immutable, and thus likely to form the basis of linguistic segmentation in both the production and perception domains.

METHODS

The "supra-moraic" prosodic factors of interest in this study were overall speaking rate and sentence-level focus. In order to manipulate these parameters, a set of nine paired sentences was compiled. In each pair, one sentence placed the target word in a position of sentence-level focus (contrastive focus), whereas the other sentence placed the target word in a neutral position (broad focus). Contrastive focus was achieved by constructing sentences that contrasted the target word with another word in the sentence. An example of a sentence pair is given below in Table 1.

Table 1. Example of a sentence pair with the target word in bold.

Contrastive focus:
Tanaka san wa umibe ***ga*** suki na no de wa naku, umibe ***mo*** suki na no de su.
'It's not that Mr. Tanaka *only* likes the seaside, it's that he ***also*** likes the seaside.'
Broad focus:
Tanaka san wa umibe ***mo*** suki na no de su.
'Mr. Tanaka ***also*** likes the seaside.'

The target words consisted of one, three, or five moras, with three tokens for each mora length. Thus, there were three tokens for each of three mora lengths for each of two focus conditions giving a total of 3x3x2=18 sentences. Two separate lists were prepared: one for the "contrastive focus" sentences and one for the "broad focus" sentences.

Four native Japanese speakers, two males (both age 32 years) and two females (age 32 and 29 years), served as subjects. All speakers were employees at ATR Research Laboratories in Kyoto, Japan, and did the recordings during regular work hours. Although all of the subjects were currently living in the Kyoto prefecture, they were all originally from the Tokyo region and were considered speakers of the Tokyo dialect.

The speakers were recorded in an anechoic chamber at ATR Human Information Processing Research Laboratories in Kyoto, Japan. Each speaker read the two randomized lists of sentences three times: first at a "normal" speaking rate, then at a fast rate, and finally at a slow rate. No attempt was made to ensure that all speakers spoke at the same rates in an absolute sense; rather, the focus of our attention was on achieving three different rates for each speaker in a relative sense. Different randomization orders were used for each subject and for each rate. The order of reading the "contrastive focus" and the "broad focus" lists was consistent across speaking rates for each subject, but was counter-balanced across subjects.

The recordings were digitized and analyzed using the Entropic Signal Processing System on a SUN SPARCstation 5 in the Speech Research Laboratory at Indiana University. The target word in each sentence was marked by time cursors in the waveform. Spectrographic displays were used in conjunction with the waveform displays to determine the onset and offset of the target word. Total word durations and peak fundamental frequencies of the target words were extracted from these labeled portions of the speech files.

RESULTS

Duration

Figure 1 shows the effect of speaking rate and sentence focus on total word duration (in milliseconds) as a function of the number of moras in the word. This plot shows the data for all four speakers pooled; however, the general pattern is the same for each individual speaker. As seen in this plot, at each of the three speaking rates there is a linear relationship between number of moras and total word duration. This plot also indicates that this relationship is not affected by whether the target word is embedded in a sentence with contrastive focus or not. A four-factor ANOVA with Speaker (four levels), Number of Moras (one, three, five), Focus (contrastive, broad), and Speaking Rate (fast, medium, slow) as factors was performed. This analysis revealed a main effect of Number of Mora ($F(2,144)=1352.8$, $p<.01$), of Speaking Rate ($F(2,144)=399.0$, $p<.01$), and of Speaker ($F(3,144)=64.0$, $p<.01$); however there is no main effect of Focus ($F(1,144)=.269$, $p=.60$). The main effect of Speaker is due to individual differences in overall speaking rate, since absolute speaking rate was not controlled across speakers and speakers differed in their "normal" speaking rate. Speakers also differed in the extent to which the fast and slow speaking rates differed from the medium rate; however, for each of the four speakers we find the same significant pattern of results that we find for the pooled data. In all cases, there is a consistent linear relationship between total word duration and number of moras in the word, and, this linearity remains unperturbed by variation in speaking rate and by variation in focus condition.

This observed linearity confirms that in Japanese the time dimension is constrained by a strict principle of mora timing. In light of the finding that variation in sentence-level focus is not reflected in the acoustic signal by durational differences, we performed a comparison of peak fundamental frequency for the target word across the two focus and three rate conditions. Our expectation was that the contrast between the target words in the contrastive-focus versus broad-focus sentences would be reflected in the acoustic signal by a difference in fundamental frequency peak (in Hertz). Such a finding would confirm that the sentence pairs we used in this study were effective in eliciting a

difference between contrastive- versus broad-focus sentences in the acoustic-phonetic domain despite the lack of a difference in duration. This would in turn validate our interpretation of the consistency of mora timing in Japanese, even under conditions which might be expected to affect the time dimension.

Fundamental frequency

Figure 2 shows the peak fundamental frequency (in Hertz) for the target words in the contrastive- and broad-focus sentences as a function of number of moras (top panel) and as a function of speaking rate (bottom panel). These plots show the data for all four subjects pooled; however, the general pattern is the same for each individual speaker. As seen in both panels of Figure 2, the target word F0 peak in the contrastive-focus condition is consistently higher than in the broad-focus condition. Furthermore, the F0 peak is not affected by overall speaking rate (bottom panel). In an ANOVA with Speaker, Focus, Number of Mora, and Rate as factors, there is a main effect of Speaker ($F(3,144)=360.5$, $p<.01$), reflecting the individual differences in fundamental frequency range, a main effect of Focus ($F(1,144)=9.9$, $p<.01$), and a main effect of Number of Moras ($F(2,144)=23.5$, $p<.01$) possibly reflecting the effect of the different segmental structures of the particular tokens. There is no main effect of Rate ($F(2,144)=1201.8$, $p=.12$). Thus, whereas total word durations remain unaffected by variation in sentence-level focus, fundamental frequency peak does vary according to this factor.

SUMMARY AND DISCUSSION

The data in this study provide additional evidence that, in Japanese, phonetic variation in the time dimension is severely constrained by the principle of mora-timing. As shown in earlier work (e.g. [2] and [3]) segment durations at the sub-moraic level exhibit compensatory lengthening and shorting in order to achieve a target total word duration. The present study extends this work by providing evidence of the consistency of mora timing in Japanese in the face of supra-moraic prosodic variation. This constrained acoustic-phonetic structure of Japanese timing is in contrast to a language such as English where the time dimension simultaneously reflects various prosodic features [4]. Additionally, this timing regularity in the acoustic domain raises questions about its usefulness for Japanese listeners. In a series of perception experiments with English and Japanese subjects, Cutler and Otake [5] investigated whether the Japanese listeners' exhibited a sensitivity to mora boundaries in a phoneme monitoring task. Their results showed different listening strategies for the Japanese and English subjects that could be traced to the importance of moras in Japanese versus the importance of syllables in English. This result, in conjunction with the acoustic data, suggests that the Japanese mora-timing principle is an example of a general linguistic timing principle that is under the speaker's control, and to which the listener is sensitive.

ACKNOWLEDGMENTS

We are grateful to Chiemi Kuroki for help with the recordings, to Reiko A. Yamada for help with the sentence lists, and to the speakers for their time. All are affiliated with ATR Human Information Processing Research Laboratories in Kyoto, Japan. This work was supported by NIDCD Training Grant DC-00012, by NIDCD Research Grant DC-00111, and by ONR N0001491J-1261 to Indiana University.

REFERENCES

[1] Beckman, M. (1982), "Segment duration and the 'mora' in Japanese," *Phonetica* , vol. 39, pp. 113-135.

[2] Port, R., Dalby, J., & O'Dell, M. (1987), "Evidence for mora timing in Japanese," *J. Acoust. Soc. Am.*, vol. 81, pp. 1574-1585.

[3] Han, M. (1994), "Acoustic manifestations of mora timing in Japanese," *Journal of the Acoustical Society of America,* vol. 96, pp. 73-82.

[4] Behne, D. and Nygaard, L. (1992) "Concurrent effects on duration I: Vowels," *Research on Speech Perception Progress Report No. 17.* Bloomington, IN: Speech Research Lab., Indiana Univ.

[5] Cutler, A. and Otake, T. (1994), "Mora or phoneme? Further evidence for language-specific listening," *Journal of Memory and Language,* vol. 33, pp. 824-844.

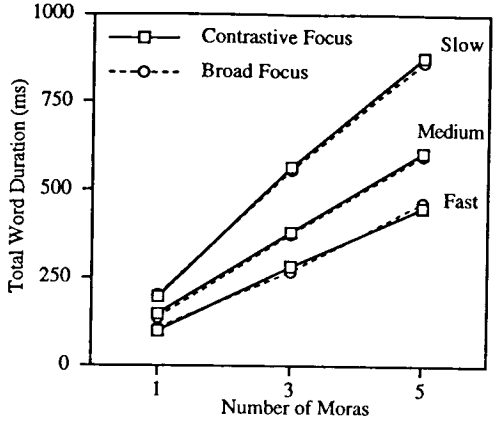


Figure 1. Total word duration as a function of number of moras for words in sentences with contrastive and broad focus at three speaking rates.

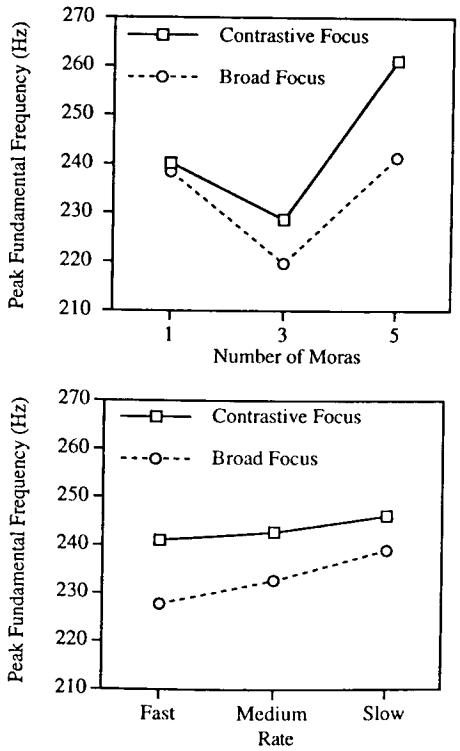


Figure 2. Peak fundamental frequency as a function of number of moras (top panel) and speaking rate (bottom panel) for words in sentences with contrastive and broad focus.

INCREMENTAL FOCUS-ACCENT-REALISATION IN SYNPHONICS

Carsten Günther, Claudia Maienborn, Andrea Schopp, Soenke Ziesche
University of Hamburg, FB Informatik, AB WSV, D-22527 Hamburg, Germany

ABSTRACT

The SYNPHONICS system represents a new, incrementally based approach towards the prosodic focus-accent realisation within the framework of a cognitively motivated concept-to-speech architecture. We want to provide an answer to the question how focus/ background calculation determines an appropriate metrical and tonal planning under recourse to semantic focus information of fragmentary increments.

1. SURVEY OF THE CSS SYNPHONICS

The Concept-to-Speech system SYNPHONICS[1] adopts a cognitive approach to a computational linguistic model of language production that combines results from psycholinguistic research about the time course of human language production with recent developments in theoretical linguistics and phonetics concerning the representation of semantic, syntactic, phonological, and phonetic-articulatory knowledge. The aim of the project consists in developing a system that covers the incremental generation of utterances from pre-linguistic conceptualisations to the formation of phonological structures, which are in turn interpreted phonetically, yielding an articulatorically specified input to a speech synthesis module. The SYNPHONICS system consists of three central processing units: a Conceptualizer, a Formulator (grammatical and phonological encoding), and an Articulator (Figure 1). Linguistic objects and rules are represented as typed feature structures in a formal specification language (ALE, Attribute Logic Engine [2]).

[1] SYNPHONICS is an acronym for Syntactic and Phonological Realization of Incrementally Generated Conceptual Structures, for a detailed description of implementational issues cf. [1].

2. INCREMENTAL COMPUTATION OF INFORMATION STRUCTURE

Among the linguistic phenomena which are analysed within the SYNPHONICS framework, emphasis is placed on investigations concerning the syntactic and prosodic realization of different information structures (e.g. focus/background structure) in accordance with conceptual and contextual variations. We argue that certain meaning distinctions triggered by changes in information structure are reflected by prosodic means without any additional support from syntax [3]. Therefore, within SYNPHONICS, a direct semantics/phonology interface is conjectured in addition to the commonly assumed syntax/semantics and syntax/phonology interfaces. This enables the phonological component to access semantic information directly. We want to provide an answer to the question how focus/ background calculation determines an appropriate metrical and tonal planning under recourse to semantic focus information of fragmentary increments.

Generally, theories of focus/ background structure and their accentual realisation consider whole sentences as the relevant domain of application. From the viewpoint of language processing, however, the sentence level is surely ruled out as primary processing unit. Incremental language production implies that the components of the language production system are enabled to process fragmentary input (so-called increments). Increments pass sequentially through succeeding components, so that each component operates in parallel on a distinct fragment of the input structure.

Under recourse to a conceptual knowledge base, the Conceptualizer of the SYNPHONICS system creates a conceptual structure CS comprising the propositional content of the planned utterance and a contextual structure CT containing the currently

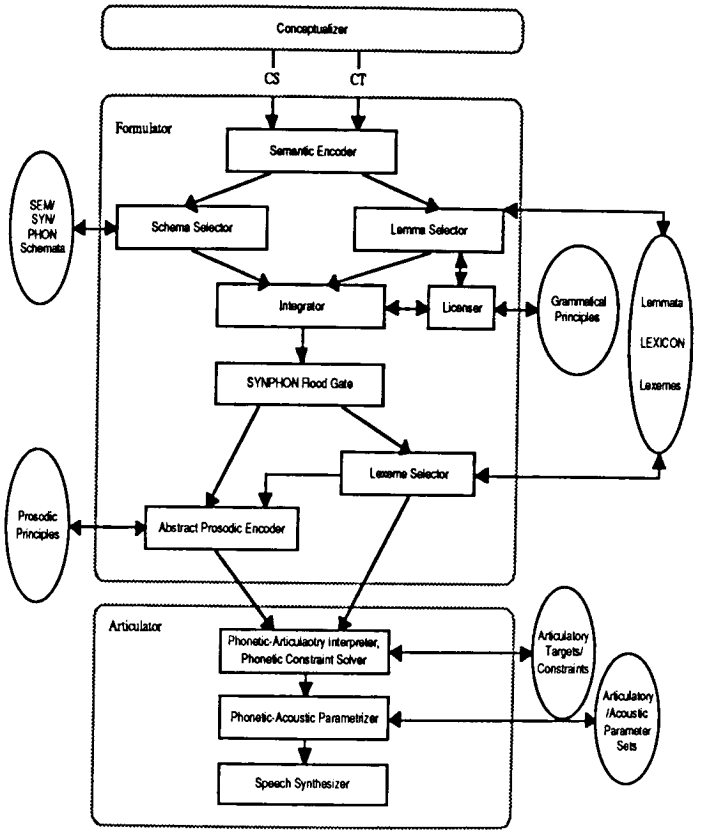


Figure 1. The SYNPHONICS architecture

relevant parts of the contextual environment. Relevant information about conceptual entities is represented in terms of referential objects (*refo*). Within the Formulator, a processing unit called Semantic Encoder generates a genuine linguistic meaning representation SEM from these input structures CS and CT.

The task of computing information structure in terms of focus/background structure is performed at the level of semantic encoding by evaluating how the informational status of the current increment fits into the focus/background structure of the whole utterance. We argue that the only information needed, besides information about the focus/background structure of the increment itself, is information whether a focused increment is part of a larger focus domain or not. A solution for the issue of this determining can be provided by the notion of context established in SYNPHONICS. The context representation can be seen as expressing the informational demand, the speaker wants to fulfil with his utterance. All contextual parameters relevant for the actual utterance are collected into the context representation CT. An adequate utterance has to meet these contextual requirements.

During semantic encoding, each increment is checked whether the information supplied fulfils the informational demand expressed by CT – in this case it belongs to the focus of the utterance – or wether it pertains to the

part of CT mutually known by speaker and hearer, i.e. the background. In the focus case, we further need to determine whether the currently processed increment fulfils the utterance's informational demand exhaustively or only partially. In the former case, we are dealing with narrow focus, in the latter the information expressed by the increment is part of a larger focus domain, i.e. we are dealing with wide focus. In case, CS directly contradicts the actual context CT, contrastive focus is assigned. The result of the computing algorithm consists in the classification of the corresponding restriction elements of the increment's semantic representation as *widely_focused, narrowly_focused, contrastively_focused,* or *non_focused* (i.e. *background*).

3. INCREMENTAL PROSODIC REALISATION OF FOCUS STRUCTURE

Differences in focus/background partitioning of semantic representation trigger different phonetic realisations by prosodic means. In German, focus type information is prosodically marked essentially by a F0-movement, the pitch accent, but also by lengthening and an intensity peak. The different focus structures cause different accent patterns. An abstract prosodic planning process interprets focus type information into an abstract prosodic feature representation (in terms of metrical pattern and accent tones) which is transformed into concrete tonal, durational and intensity parameters. Dealing with incrementallity at the processing stage of focus realisation, the complete Focus Domain (even in the case of wide focus) has not necessarily to be exhaustively specified and only partial syntactic tree structures are accessible for accentual planning.

In the following, focus realisation rules are presented that cover the determination of prominence degrees of constituents in the case of different Focus Domain sizes. These rules are variants of one general focus realisation rule and refer to different structural conditions. Due to lack of space, we neglect the prosodic interpretation of narrow and contrast focus and sketch solely the accent realisation of a wide Focus Domain. The realisation of wide focus turns out to be a more intricate problem since global structural knowledge has to be taken into account at large. The complete Focus Domain is expressed by one nuclear accent, but the exact accent placement depends on semantic and syntactic conditions. Generally speaking, in case of wide focus on VP level (focusing of the verb and its complements and adjuncts), focus is realised on the verb adjacent complement or, in case of the occurrence of a verb adjacent adjunct, on the verb itself. Preceding constituents (either complements or adjuncts) of the Focus Domain carry phrasal accent (secondary stress). Thus, a focus-accent mapping that proceeds incrementally has to check whether the current increment is situated in a verb adjacent position or not. Therefore, in (1), the non-verb-adjacent argument *Peter* is assigned phrasal accent.

(1) Maria hat [PETER das BUCH gegeben]$_F$.

(Mary has given the book to Peter.)

The rule in Figure 2 licenses the assignment of phrasal accent (*phras_acc*) to widely focused non-verb-adjacent complements. In case that a verb-adjacent constituent is selected at the Syntax-Phonology interface it has to be checked whether this constituent is a complement or an adjunct. Example (1) illustrates that verb-adjacent complements (*das Buch*) carry nuclear accent. The rule in Figure 3 (usually named *Focus Projection Rule*)

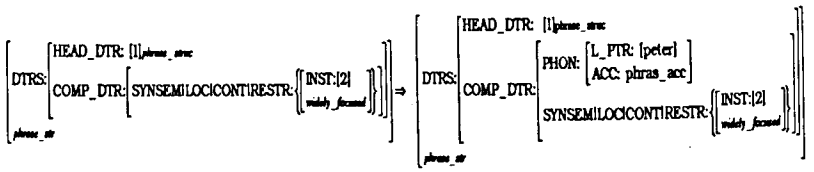


Figure 2. Phrasal-Accent Rule for non-verb-adjacent complements

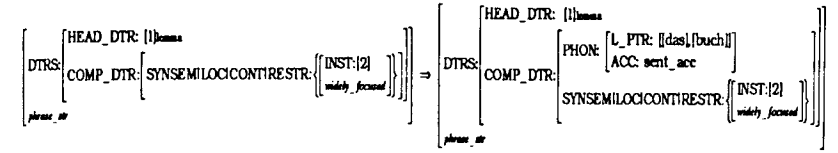


Figure 3. Sentence-Accent Rule for verb-adjacent complements

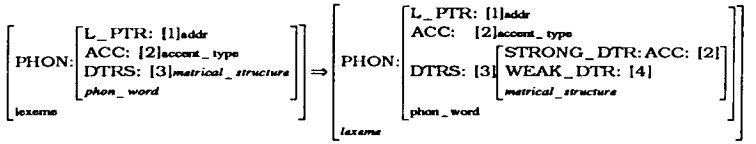


Figure 4. Accent Percolation Rule

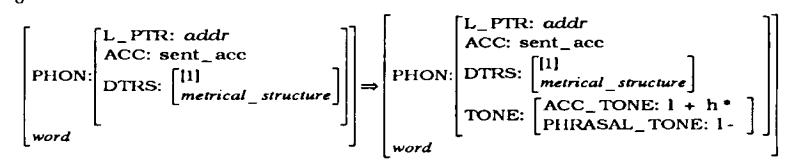


Figure 5. Tonal Sentence Accent Rule

determines the verb-adjacent focus exponent and assigns the sentence accent.

In case of widely focused verb-adjacent adjuncts (cf. 2), sentence accent is realised on the verbal head. This is ensured by a special rule set that assigns phrasal accent to the adjunct and sentence accent to the verb.

(2) Maria hat [nachts GESCHLAFEN]$_F$

(Mary has slept at night.)

The subsequent prosodic encoding and its phonetic interpretation (cf. Figure 1) in terms of articulatory and acoustic parameter settings ensures the appropriate acoustic realisation of information structuring. E.g., the application of the Accent Percolation Rule (Figure 4) licenses the realisation of the sentence or phrasal accent on the word accent bearing syllable, the designated terminal element (DTE). Prosodic rules operate on structures of the abstract data type *metrical tree*.

On the designated syllable, the prenuclear and nuclear accent is realised as bitonal (e.g. sentence accent: L+H*; contrastive accent: L*+H) (Figure 5) or monotonal (phrasal accent: H*) accent tones. The assignment of the phrasal tone L- reflects the end of a Focus Domain.

Phonetic-acoustic interpolation rules subsequently parametrise the F0-contour on the accent-tone adjacent syllables. For synthesizing speech, a Klatt-based *Formant Syntesizer* algorithm (TU Dresden) is applied.

ACKNOWLEDGEMENT

The research reported in this paper is carried out in a research project which is funded by the German Science Foundation (DFG) within the research program of Cognitive Linguistics under grant no. Ha 1237/4-3.

REFERENCES

[1] Günther, C. (ed.,1994), *Hamburger Arbeitspapiere zur Sprachproduktion VI*, University of Hamburg.

[2] Carpenter, B. (1992): *The Logic of Typed Feature Structures*. Cambridge: Cambridge Univ. Press.

[3] Günther, C., C. Maienborn, A. Schopp (1994): *The processing of information structure in SYNPHONICS*. In: P.Bosch & R. van der Sandt (eds.): *Proc. of the interdisciplinary conference of the 10th anniversary of the journal of semantics "Focus & natural language processing"*. Schloß Wolfsbrunnen.

CONTRASTIVE EMPHASIS IN ELICITED DIALOGUE: DURATIONAL COMPENSATION

D. Erickson and I. Lehiste
The Ohio State University, Columbus, Ohio, USA

ABSTRACT

Duration of words in an utterance perceived by listeners as emphasized are longer than the counterpart words in reference utterances (spoken without emphasis); moreover, the non-emphasized words in utterances with emphasis are shorter than the counterpart words in the reference utterances. Thus, emphasis involves temporal rearrangement of all the words in an utterance, not just the word receiving emphasis.

INTRODUCTION

It is generally known that words produced with contrastive emphasis frequently exhibit increases in duration, intensity, and F0 level, e.g., [1]. Here we study the temporal structure of utterances that differ in the presence or absence of contrastive emphasis on one of the words in otherwise identical sentences. The utterances consist of responses involving a 3-digit sequence with the words "five" or "nine" followed by "Pine Street". The utterances were elicited in a dialogue format designed to have the speaker repeat the same correction up to five or six times [2].

Our hypothesis is that emphasis is a phrase level phenomenon; the emphasized digit will be longer in duration relative to the other digits in the sequence; it will also be longer than the corresponding digit in the utterance spoken with no emphasis (hereafter referred to as the "reference utterance" as opposed to the "emphasis utterance"). Since the duration of words in English varies according to their position in the utterance (i.e., phrase final words are longer in duration than the other words), we hypothesize that the percentage increase needed for a word to be perceived as emphasized will also vary.

METHODS

Approximately 38 to 70 target utterances were elicited from each of four speakers of American English in an experimental paradigm that called for contrastive emphasis on one of three digits, and approximately 12 to 18 target utterances intended to be produced as reference utterances. The target utterances were of the type "595 Pine Street", "559 Pine Street", and "959 Pine Street."

The speakers were instructed by the first author to pretend that this was a telephone conversation and to reply to the questions by reading the prompt on the monitor. If the elicitor indicated she was having problems hearing the response clearly, the speakers were asked to "not read the prompt in the monitor screen but to try to get the correct information across according to what the monitor specified." The elicitor sat out of sight but within hearing distance of the speaker. For a subset of responses, the elicitor deliberately misunderstood the speaker's answer repeating the digit sequence with the initial, medial or final digit incorrect. The speaker responded by giving the correct information without reading the monitor prompt. Sometimes the elicitor asked the speaker to repeat the correct digit sequence five or more times. We refer to the series of exchanges between elicitor and speaker as a "dialogue set"; it always included one reference utterance and several repetitions of the utterance with the corrected information. A typical dialogue with one speaker is given below. The answer by the speaker to the first question is referred to as the "reference utterance" and is indicated in italics below.

Dialogue 2 (S4)

1.DE: Where do you live?
S4: *I live at 595 Pine Street.*
2.DE: I'm sorry, that was 599 Pine Street?
S4: No, 595 Pine Street.
3. DE: I'm still not getting it. 599 Pine Street?
S4: I live at 595 Pine Street.
4.DE: You're saying, 599 Pine Street?
S4: No, 595.
5. DE: 599 Pine Street, right?
S4: No, 595 Pine Street.

It was assumed that the speaker would produce the target utterances first with no contrastive emphasis, and then in response to the elicitor's misunderstanding of one of the digits in the utterance, with emphasis on one of the digits. We found, however, that it was not always obvious which was the emphasized digit. We ran formal perception tests with 20 listeners and 2 randomizations of the target utterances. The target utterances were the three-digit phrase plus "pine street" which had been extracted from the speaker's response. (Occasionally the speaker would not say "pine street", only the three digit sequence; in which case, only the three-digit sequence was used.) A separate test was made for each of the four speakers in the data base. The listeners' task was to indicate which digit the speaker was making a correction on, and to guess if they were not sure. The results of the perception test indicate that not all of the utterances intended to contain an emphasized digit were identified as such by listeners: only 46% to 70% of possible instances were heard by listeners as carrying contrastive emphasis. We also found that although generally the 3-digit sequence was spoken as part of a single phrase, some of the 3-digit sequences were spoken with each digit as a separate phrase.

Using the Waves+ software, the acoustic signal was digitized and the durations of initial, medial and final digits in the target utterances were measured. In this analysis, we excluded those utterances that were spoken without "Pine Street" or that had phrase breaks between the digits. (Because S1 tended to produce the 3-digit sequence without "Pine Street", to insert phrase breaks between the digits, and generally to produce utterances that were not well perceived by listeners as having emphasis on the intended digit, her data are not analyzed here.)

RESULTS AND DISCUSSION

The durations of "5" and "9" were measured from the reference utterances of the three speakers. These durations did not vary as much as a function of their identity as they did as a function of the position of the digit in the utterance; thus, we averaged together the initial "5's" and "9's", the middle "5's" and "9's", and the final "5's" and "9's".

We measured by position in the phrase the durations of the digits in the reference utterance and those in that particular utterance within the dialogue set that were best perceived by listeners to have emphasis on the intended digit. Figure 1 shows results for one of the speakers. The emphasized digit is always longer in duration than the other digits in the 3-digit sequence, which observation is compatible with findings from other studies of acoustic duration.

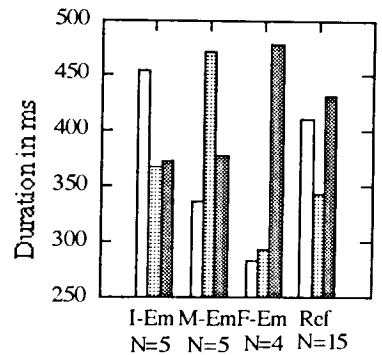


Figure 1. Durations of initial (white), middle (gray), and final (dark) digits for the utterances when the listeners reached the highest agreement on the digit intended to be emphasized in initial, middle and final position, and for the reference utterances.

Moreover, the emphasized digit is longer in duration than the corresponding digit in the utterance spoken with no emphasis. The unemphasized digits are shorter in duration than their counterparts in the reference utterances. For instance, the duration of the final digit in the utterances with final emphasis is clearly longer than the final digit in the reference phrase, and the durations of the initial and middle digits in utterances with the final digit emphasized are decidedly shorter than the initial and middle digits in the reference phrase. This same pattern is seen for the other two speakers (not shown here.)

In order to compare the distribution of durations among the 3 digits across the

different speakers, the durations were calculated in terms of percentages of the total duration of the 3-digit sequence.

Figure 2 shows the results for speaker 3. Note there is a progression in amount of duration needed for a digit to be perceived as emphasized as a function of the position of the digit in the sentence, with the emphasized initial digits constituting 38% of the total duration, middle digits, 40% of the total duration, and final digits, 45% of the total duration.

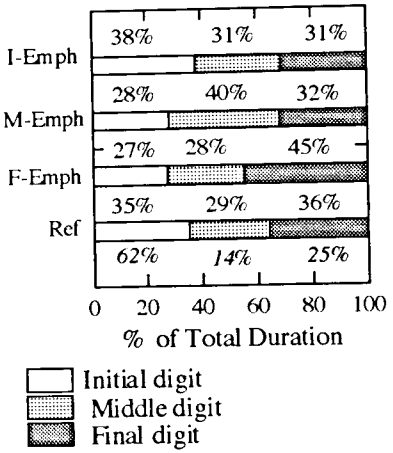


Figure 2. Percent total duration of digits displayed in Figure 1. Percentages are given above each horizontal row; the percentages in italics below the reference bar graph indicate how listeners judged emphasis on the reference utterances.

The bottom row in the graph shows the reference utterances with a break-down of 35%-29%-36% on the initial, middle and final digit respectively. The numbers in italics below the reference utterance indicate how listeners judged emphasis on the reference utterances, given a forced choice task. 62% of the time, listeners assigned emphasis to the initial digit (even though the final digit was longer in duration); 14% of the time, they assigned emphasis to the middle digit, and 25% of the time, to the final digit. It is curious that the duration of the initial digit constitutes 35% of the total duration, yet 62% of the time was heard by listeners as emphasized. Other acoustic cues of intensity and F0 also influence the perception of emphasis. Measurements of peak intensity (in rms) for each digit for this speaker show relatively greater intensity on the initial digit than on the other digits of the reference utterance; this must also be affecting the listeners perception of emphasis. For the other two speakers also, listeners consistently assigned emphasis to the initial digit , even though the initial digit made up approximately only one third of the total duration. Intensity and F0 measurements remain to be made for these speakers.

Table 1 compares the percent of total duration of the 3-digit sequences in the emphasized utterances with those in the reference utterances for each of the three speakers. All three speakers show strikingly similar patterns of duration: the emphasized digit is always greater than its unemphasized counterpart in the reference utterance, and the other digits in the emphasized utterance are always shorter than their counterparts in the reference utterance. The one exception to this is the duration of the middle digit of the utterance with the initial digit emphasized (for speaker 3), which is slightly larger (2%) than the middle digit in the reference utterance.

There is a tendency, especially for speakers 2 and 3, for the initial digit to require less of an increase in duration compared to the middle or final digits in order for it to be heard as an emphasized digit. We wondered why this might be. It may be that the initial emphasized digit increased in duration only by 1% - 3% (for speakers 2 and 3) because when the initial digit made up approximately one third of the total duration of the reference utterance, it was heard as emphasized by over 50% already of the listeners. Thus, only a slight increase in duration would be needed for the initial digit to be heard as emphasized.

Also, it seems that, at least for speakers 2 and 3, a greater increase in duration is required for the middle or final digit to be heard as emphasized than the initial digit.

In summary, it seems that emphasis involves rearrangement of the durational relationships within the utterance, not just an increase in duration of the emphasized item. The emphasized item is increased, and duration is taken off from the other items; the amount of increase/ decrease varies according to the position of the word in the phrase. We suggest that by lengthening the emphasized word and shortening the other words within the utterance, the speaker maximally differentiates the utterance, thus increasing the chances that emphasis will be perceived by listeners.

Table 1. Comparison of the percentage of a digit constituted of the total duration of the 3-digit sequence in reference utterances and utterances with an emphasized digit. Data are shown for speakers S2, S3, and S4.

S2	Initial	Middle	Final	Initial	Middle	Final
Reference (N=8)	34%	32%	34%			
Initial Emphasis (N=1)	35%	31%	34%	**+1%**	-1%	0%
Middle Emphasis (N=3)	29%	41%	30%	-5%	**+9%**	-4%
Final Emphasis (N=4)	27%	28%	45%	-7%	-4%	**+11%**

S3	Initial	Middle	Final	Initial	Middle	Final
Reference (N=15)	35%	29%	36%			
Initial Emphasis (N=5)	38%	31%	31%	**+3%**	+ 2%	- 5%
Middle Emphasis (N=5)	28%	40%	32%	-7%	**+11%**	-4%
Final Emphasis (N=4)	27%	28%	45%	-8%	-1%	**+9%**

S4	Initial	Middle	Final	Initial	Middle	Final
Reference (N=17)	33%	32%	35%			
Initial Emphasis (N=8)	39%	30%	31%	**+6%**	- 2%	- 4%
Middle Emphasis (N=5)	32%	37%	31%	-1%	**+5%**	-4%
Final Emphasis (N=4)	30%	29%	41%	-3%	-3%	**+6%**

References

[1] Lehiste, I.(1970). Suprasegmentals. Cambridge, MA: MIT Press.

[2]Erickson, D., Lenzo, K., & Sawada, M. (1994). Manifestations of contrastive emphasis in jaw movement in dialogue. Proc. International Conference of Spoken Language Processing, Yokohama, Sept. 1994.

Acknowledgments

This work has been supported in part by a research fund from ATR, International, given to O. Fujimura.

THE ROLE OF STRESS AND ACCENT IN THE PERCEPTION OF RHYTHM

Cynthia Grover and Jacques Terken
Institute for Perception Research (IPO), Eindhoven, The Netherlands

ABSTRACT

We ask whether stress or accent alone, or in combination, induce the perception of rhythm. According to Grønnum [1], perceiving speech rhythm depends on the distribution of accented syllables, and durational variation is unimportant. We report an experiment on perceiving speech rhythm whose outcome confirms the primary role of accent in determining rhythmicity.

INTRODUCTION

Phoneme durations are highly variable [2], and so it seems unlikely that their duration is tightly controlled. Perhaps their durational variation is constrained by higher level influences, as our own recent work on German suggests [3]. One such influence could be speech rhythm.

The definition of speech rhythm is problematic. The usual implication of the words is regular alternation of two sorts of units, one more prominent and (at least) one less so. Often, this contrast is labelled in terms such as strong versus weak, eg.[4], stressed versus unstressed, eg.[5], or accented versus unaccented, eg.[1]. We use 'accent' to mean 'pitch accent', and stress to mean 'without pitch accent and bearing lexical stress'. Our purpose is to define rhythmicity, and so to provide a better basis for discussing rhythmic constraints in speech.

We also wish to gather information about the sources of the perception of rhythm. Specifically, to what extent do stress and accent induce rhythmicity, or the perception that speech is rhythmic?

We suppose that listeners might consider stimuli that can be analysed into regular metrical linguistic units (ie. feet) to be rhythmic. We ask whether this foot structure is defined more clearly by accents or stresses.

Our hypotheses are as follows. Stresses and accents are assumed to serve as the heads of feet.

1. A uniform foot structure (always the same number of unstressed syllables per foot) and uniform head type (always accent or always stress) should give rise to a clear perception of rhythmicity.
2. Variation in the foot structure or head type within the phrase should induce less rhythmicity than when these are regular. This variation could take two forms: A. The number of unstressed syllables per foot varies, or B. the type of head varies (mixed accent and stress).
3. Variation in both foot structure and head type should induce the least rhythmicity.
4. We assume that subjects initially hypothesize foot structure based on the initial part of a stimulus. We therefore propose that inconsistent foot structure early in a stimulus is more salient than variation late in a stimulus, so early variation should be less rhythmic.

METHOD

We wanted to know how rhythmicity related to perceived foot structure and head type, so we had subjects perform 3 tasks on reiterant "mamama.." stimuli: A. rating their rhythmicity, B. picking prominences, and C: placing boundaries.

Subjects

Eighteen Dutch adults participated in two half-hour experimental sessions.

Materials

A male native Dutch speaker produced series of the syllable "ma" with pitch accents and stresses occurring at regular intervals (ie. every third or fourth syllable). We selected a typical exemplar each of accented, stressed, and unstressed syllables, and concatenated them to produce 8 and 9 syllable strings of reiterant "mamama.."'s. Table 1 shows the syllable types' specifications.

Table 1. Syllable (ma) specifications. D: duration (ms); F_0: Hz; A: amplitude (N/m^2).

	D	F_0max	F_0min	Amax	Amin
Accent	269	85	126	12830	-12910
+Stress	216	85	89	4991	-4402
-Stress	197	85	88	2935	-3565

Nine full "mamama.." strings were concatenated, and from these an additional 9 truncated strings were made by removing the last syllable of the string. The patterns are given in Table 2.

Procedure

Dutch subjects performed 3 tasks: rhythmicity rating, prominence picking, and grouping in each of two sessions. They heard the stimuli through headphones. Each task began with 3 practice stimuli. Then for each of the 3 tasks, the 18 strings were called up by a control program. Subjects had to complete the indicated task for each string before hearing the next string.

Subjects rated rhythmicity on a scale of 1 to 10 (10: extremely rhythmic). They could pick as many prominences or boundaries as desired by clicking a mouse to select syllables or boundaries (represented graphically on screen).

Analysis

We performed repeated measures multiple regressions. For every syllable the expected response ("prominent" for stresses or accents, and "not prominent" for unstressed syllables) was registered in the datafile for matching against the actual subject response.

RESULTS AND DISCUSSION

Before all analyses, we removed the variance due to control variables: individual differences and the order of presentation of the stimuli.

Rhythmicity

As predicted, there was a significant difference in the ratings for the different accent patterns, $F_{(8,621)}=26$, $p<.01$. Patterns with consistent head type or foot structure or both (strings 1 to 8 (S1-8)) were judged to be more rhythmic than those with variable foot structure and head type (S9), $F_{(8,621)}=23.5$, $p<.01$ (see Table 2). However, the clearest division between ratings fell between strings 1-6 versus S7-9, $F_{(1,621)}= 160$, $p<.01$. Thus an important factor in rhythmicity is not simply the presence or absence of consistent foot structure, but the particular type of inconsistency. S2 and 4 can be analyzed as alternating, except that the place of the third head is filled by a syllable which is not prominent; S7 and 8 cannot be analyzed as alternating.

Early inconsistency in foot structure (S8) induced less rhythmicity than the late inconsistency in foot structure (S7), but the difference failed to reach statistical significance, $F_{(1,621)}=3.5$, $p<.05$.

The strings with consistent foot structure and head type (S1 and 3) were rated as more rhythmic than the strings with just the consistent head type (S2 and 4), $F_{(1,621)}= 32.7$, $p<.01$, and also as significantly more rhythmic than the strings with just consistent foot structure (S5 and 6), $F_{(1,621)}= 25.3$, $p<.01$. Thus, our hypotheses that rhythmicity depends upon the consistency of foot structure and head type are supported.

Further, it is clear that a head may be defined by pitch accent or stress. While the ratings were higher for the strings

Table 2. Rhythmicity, prominence and grouping per string. The numbers of boundaries and of prominent syllables are given. (#:string number. M:mean rhythm rating; sd: standard deviation; ma: unstressed; MA: stressed; MA: accented; P: number of prominences; B: number of boundaries. Parenthesis shows line of truncation.)

#	Rhythm M	sd	Pattern / 2a4a6a8a
1	7.7	2.0	maMAmaMAmaMAmaMA(ma
			P:4 66 4 66 4 66 6 63 1
			B:13 45 22 45 22 44 20 11
			2a4a6u8a
2	6.3	1.9	maMAmaMAmamamaMA(ma
			P:5 67 5 66 7 6 12 56 1
			B:12 40 16 47 15 6 22 7
			2s4s6s8s
3	7.2	2.6	maMAmaMAmaMAmaMA(ma
			P:8 28 8 25 8 25 8 24 5
			B:13 22 15 23 15 21 15 7
			2s4s6u8s
4	6.2	2.2	maMAmaMAmamamaMA(ma
			P:12 35 12 30 9 8 7 40 4
			B: 7 19 11 33 13 12 22 2
			2a4a6s8a
5	6.5	1.8	maMAmaMAmaMAmaMA(ma
			P:4 64 7 61 9 8 12 49 2
			B:14 36 22 46 21 7 20 5
			2s4s6a8s
6	6.3	2.1	maMAmaMAmaMAmaMA(ma
			P:5 12 4 11 7 58 7 4 1
			B: 6 9 6 20 34 25 9 1
			2a4a7a8a
7	5.0	2.2	maMAmaMAmamaMAMA(ma
			P:4 67 4 63 7 11 60 55 3
			B:10 37 20 43 15 29 21 6
			2a3a6a8a
8	4.4	2.2	maMAMAmamaMAmaMA(ma
			P:4 67 68 8 12 54 20 41 2
			B:13 29 43 17 22 17 23 3
			2s3a5s8a
9	5.1	2.0	maMAMAmaMAmamaMA(ma
			P:4 6 62 7 5 7 6 55 3
			B: 1 16 41 19 7 8 19 3

with an accent (S1 and 2) rather than a stress head (S3 and 4), this difference was not statistically significant ($F_{(1,621)}$= 2.2, p<.05).

Prominence Picking

Accented syllables were perceived as prominent more often than were stressed syllables, p<.01. There was no statistically significant difference in the proportion of matches of response to predicted outcome for the accented and unstressed syllables (84% for accents, 90% for unstressed syllables, and 27% for stressed syllables).

The stimulus type accounted for over .08 of the variance in the match of predicted to actual prominence, p<.001. Subjects picked prominences in the strings with either consistent head type or foot structure (S1-8) more in line with our predictions than in the string with variable head type and foot structure (S9), p<.01. There was virtually no difference in the proportion of matches to predicted prominences along the lines that were important in judging rhythmicity, namely between S1-6 and 7-9. Thus rhythmicity and prominence are distinct percepts.

Prominences were picked largely as predicted wherever accents consistently served as foot heads; the worst match of predicted to actual prominence occurred in S6 and 9, where both accent and stress were present. There were fewer matches toward the end of the truncated strings, p<.01.

Grouping

The number of groups perceived differed with stimulus type, $F_{(8,620)}$= 18.6, p<0.05. We had thought that subjects might attempt to analyze strings into feet to increase their rhythmicity. This probably does not occur. In the string with variable foot structure and head type (S9) subjects nominated fewer groups than for other strings, $F_{(8,620)}$= 24, p<.01. In the other 2 strings with low rhythmicity (S7 and 8), subjects nominated many groups, but the ratings were nonetheless low. Also, the contrast between strings which was important in judging rhythmicity (S1-6 vs S7-9) was not important in the grouping data. Thus grouping and rhythmicity judgment are distinct percepts.

The number of groups shows a fair correlation with the number of prominences, r =.51, $F_{(1,626)}$ = 186, p<.01. Subjects placed boundaries consistently next to accents and stresses, generally preferring to end a group with a stress or accent. The traditional view of starting feet with a stress is then not upheld.

Accent induces division into groups more reliably than does stress; strings of feet headed by stresses contain significantly fewer groups than those with feet headed by pitch accents (S3,4 and 6 vs S1,2 and 5: $F_{(1,620)}$= 91, p<.01).

Lastly, longer stimuli contained more boundaries. This could mean either that strings ending in a stress or accent are broken into groups differently when they end in an unstressed syllable, or that listeners tend to break strings of any length into units of a consistent size, in which case one would expect more groups in longer strings.

CONCLUSION

Our expectations about rhythmicity were in general upheld. Subjects distinguished degrees of rhythmicity clearly. High rhythmicity of strings arose in strings with consistent foot structure and head type. Inconsistency of foot structure does not necessarily reduce rhythmicity; strings in which feet contained 1 or 3 unstressed syllables (S2 and 4) were still perceived to be quite rhythmic. Indeed, the type of inconsistency is crucial: feet with 1, 2 or 3 unstressed syllables (S7 and 8) were perceived to be less rhythmic. This difference is nicely predicted by the clock formulation of Povel and Essen [6, 7]. Our results suggest strongly that rhythmicity in speech is a function of the regularity of a unit's recurrence, which is also the basis of their clock formulation. Here regularity means the consistency of foot structure, which is confounded with consistency of duration.

In any case, in judging rhythmicity subjects do not base their decisions wholly on the number of groups perceived or of prominences picked. Prominence picking can specify beats (in Povel and Essen's terms) and grouping can specify meter. Their regularity contributes to rhythmicity. Prominences and groups are picked on similar bases, that is, with respect to predicted accent position.

REFERENCES

[1] Grønnum, N. (1993). Rhythm in regional variants of standard Danish. *ESCA Working Papers, 41*, 20-23.

[2] Campbell, W.N. (1992). *Multi-level Speech Timing Control.* Ph.d. thesis. University of Sussex, Brighton, U.K.

[3] Grover, C. & Terken, J. (1994). Rhythmic constraints in durational control. *Proceedings of the International Conference on Spoken Language Processing (Yokohama, Japan), 1*, 363-366.

[4] Giegerich, H. (1985). *Metrical Phonology and Phonological Structure.* Cambridge U. Press: Cambridge, U.K.

[5] Bruce, G. (1987). On the phonology and phonetics of rhythm: Evidence from Swedish. *Phonologica (1984); Proceedings of the Fifth International Phonology Meeting (Eisenstadt 1984)*, 21-31.

[6] Povel, D-J. (1984). A theoretical framework for rhythm perception. *Psychological Research, 45*, 315-337.

[7] Povel, D-J. & Essens, P. (1985). Perception of temporal patterns. *Music Perception, 2*, 411-440.

THE TWO-MORA FOOT IN JAPANESE
-*TANKA* RECITATION BY THE REIZEI FAMILY-

Yayoi Homma
Osaka Gakuin University, Osaka, Japan

ABSTRACT

Existence of a foot consisting of two morae in Japanese has been suggested by several linguists. However, it is very difficult to prove it phonetically in modern Japanese. I examined the very slow *tanka* poetry recitation orally handed down in the Reizei Family since the twelfth century, and found that there were two-mora units with the second mora prolonged.

1. INTRODUCTION

The basic rhythmic unit in Japanese is the mora. The mora coexists with larger units such as a word in prose [1], a poetic line [2], and a group of lines in *tanka* poems [3]. Although existence of a foot consisting of two morae in Japanese has been suggested by several linguists, it is not easy to prove its existence phonetically. Bekku[4] claimed that the rhythm of Japanese is made of quadruple time, and one beat of this quadruple time is a two-mora foot (p. 52). But his explanation has no experimental support. Poser [5] introduced Teranishi's experiment [6]. According to his experiment, "as the tempo decreased odd-numbered morae (counting from the beginning of the word) changed little in duration, while even-numbered morae lengthened considerably" (Poser [5], p. 80.) Another experiment is necessary to examine these findings in detail. The purpose of this paper is to investigate the very slow *tanka* recitation by the Reizei Family, and try to find evidence of a foot in Japanese which modern Japanese may have lost.

2. EXPERIMENT

2.1. Methods

I used a copy of a tape of *tanka* recitation recorded by Mrs. Fumiko Reizei and her father, the late Count Tametsugi Reizei. The Reizeis are called the "Family of Poetry," because they have produced great *tanka* poets and preserved precious documents and ceremonies including *tanka* recitation for eight hundred years.

Tanka were recited at poetry contests where court poets competed publicly on prescribed topics. The recitation was called "*Hikoo*." Minegishi [7] wrote that the poetry contest had its golden age in 880-1230 A.D., from the *Heian* to *Kamakura* Periods. *Hikoo* is still adopted at the annual *tanka* competition held at the beginning of the year at the Imperial Palace, and also at the Reizeis' *tanka* competition parties held four times every year on special occasions.

There are three styles of recitation. One is prose style reading, and the others are song styles with two different melodies. I have chosen the prose style without melody.

The recorded *tanka* was from the *Kokinshu, Collection of Poems Ancient and Modern*, the first Imperial Anthology (c. 905 A.D.). This *tanka* is the original of the Japanese national anthem. It is composed of thirty-two morae, in 5-7-6-7-7 morae lines.

Waga kimi wa	(5)
Chiyo ni yachiyo ni	(7)
Sazareishi no	(6)
Iwao to narite	(7)
Koke no musu made	(7)

(May our friend endure,
A thousand, eight thousand ages:
Till the smallest pebble grows
To a boulder etched with moss[8].)

2.2. Measurements

Wide-band spectrograms of the prose style readings by Mrs. Reizei (F.R.), her father (T.R.), and myself (Y.H.) were made with a Kay-Sonagraph (5500). I measured the duration of each segment, mora, foot, line, and pause.

2.3. Results

2.3.1. Line duration

Figure 1 shows the comparison of the duration (ms)of the five lines read by the three speakers. The line duration

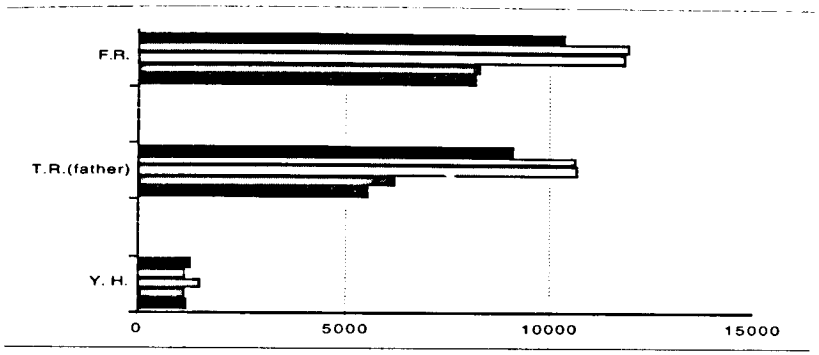


Figure 1. Comparison of the line duration(ms) of the three speakers.

includes pauses except for the last line.

The pause after each line was considerably long, especially in the Reizeis, as seen in Figure 2 on the next page.

The tempo was very slow; more than seven or eight times longer than my reading of the same *tanka*. The respective durations of the poem by F.R., T.R., and Y.H. were 50,484 ms, 42,095 ms, and 6,065 ms. The duration of lines including pauses was quite equidistant for Y.H., like my previous experiments in modern Japanese prose style reading[3], and not so much for F.R. and T.R.

2.3.2. Mora and foot duration

(1) The Reizeis' style of reading

Table 1. Mora duration (ms) of F.R.'s first line.

	1	2	3	4	5
	wa	*ga*	*ki*	*mi*	*wa*
Mora	250	775	420	680	4560
Foot	1025		1100		4560

If a word is composed of two morae as in Table 1, even-numbered morae were clearly longer than odd-numbered ones. Figure 3 illustrates F.R.'s first line.

However, if a word is composed of five morae as in *sazareishi* in Table 2, there are two ways of foot formation.

Table 2. Mora duration of F.R.'s third line.

	1	2	3	4	5	6
	sa	*za*	*re*	*i*	*shi*	*no*
Mora	250	810	515	435	890	3770
Foot(A)	1060		950		4660	
Foot(B)	1060		515	1325		3770

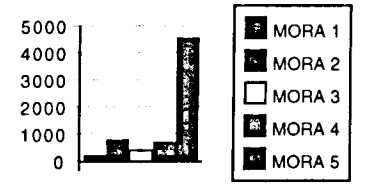


Figure 3. The mora duration (ms) of F.R.'s first line.

In Foot (A), the second mora of the second foot "*i*" is shorter than the first mora "*re*." *Sazareishi* (a pebble) is a compound noun, and there is a morpheme boundary between "*re*" and "*i*". In Foot (B), the second foot has only one mora, and therefore, is very short. However, syntactically and semantically we prefer Foot (B). This means that foot formation respects syntax,or meaning, and the two-mora foot is easily collapsed by syntactic boundaries, as seen in Figure 4 of F.R.'s third line.

Line 4 has a three-mora word "*iwao*" (a boulder). But as the next "*to*" (to) is a postposition, *iwao to* is naturally divided

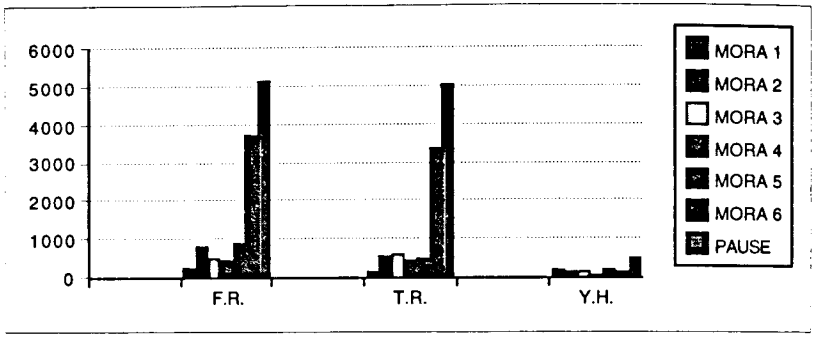


Figure 2 .Comparison of the mora duration of the third line.

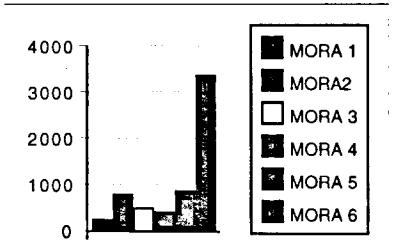


Figure 4. The mora duration (ms) of F.R.'s third line.

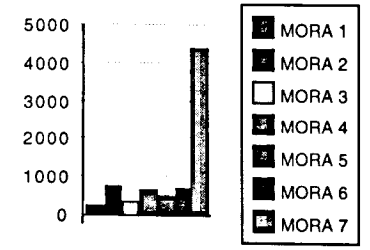


Figure 5. The mora duration (ms) of F.R.'s fourth line.

into two feet of *iwa* and *oto* as in Figure 5 of F.R.'s fourth line.

(2) Modern style of reading

Table 3 is Y.H.'s mora and foot duration of the third line.

Table 3. Mora and foot duration (ms) of Y.H.'s third line.

	1	2	3	4	5	6
	sa	*za*	*re*	*i*	*shi*	*no*
mora	190	175	165	55	215	160
Foot(A)	365		220		375	
Foot(B)	365		165	270		160

In the duration of the first foot *saza*, the second mora *za* is shorter than the first mora *sa*. The mora in modern Japanese is said to be an abstract isochronous unit of timing [1], and so is the foot in modern Japanese. The foot may be abstract, too. Figure 6 illustrates Table 3.

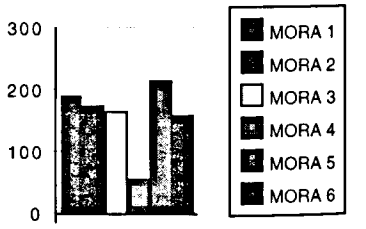


Figure 6. The mora duration (ms) of Y.H.'s third line.

In my previous experiment in modern Japanese [9], the second mora was also shorter than the first mora, as seen in Table 4 and Figure 7.

Table 4. *Mora duration (ms) of Y.H.'s seven-mora line*[9].

	1	2	3	4	5	6	7
	ha	na	no	chi	ru	ra	n
Mora	131	116	132	163	116	133	83
Foot	247		132	279		216	

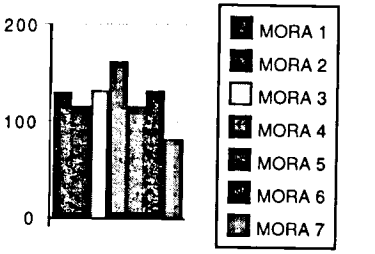


Figure 7. The mora duration (ms) of Y.H.'s seven-mora line[9].

3. CONCLUSION

The results revealed the following points in the Reizeis' recitation:

(1) The tempo was very slow; more than seven or eight times longer than my reading.

(2) The duration of the lines including pauses was, by and large, kept equidistant, but not as equal as in the modern prose style reading.

(3) The second mora was longer than the first mora, if there was no syntactic boundary between them.

(4) Extremely long prepausal lengthening was observed in the last vowel of each line.

In modern Japanese, to find acoustic evidence for foot structure is difficult. Like the mora, the foot might be an abstract isochronous unit of timing, not appearing on the surface. However, in the very slow *tanka* recitation by the Reizeis, there were two-mora units with the second mora prolonged, although they were easily collapsed by syntactic boundaries. This might be a piece of evidence for a two-mora foot in Japanese.

ACKNOWLEDGEMENT

I am grateful to Mrs. Fumiko Reizei for letting me copy the precious tape.

REFERENCES

[1] Port, R.F., J. Dalby, and M. O'Dell (1987), "Evidence for mora timing in Japanese," *Journal of the Acoustical Society of America*, vol. 81, pp.1574-1585.

[2] Lehiste, I. (1990), "Phonetic investigation of metrical structure in orally produced poetry," *Journal of Phonetics*, vol. 18, pp. 123- 133.

[3]Homma,Y.(1991), "The rhythm of *tanka*, short Japanese poems: read in prose style and contest style," *Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, France*, vol. 2, pp. 314-317.

[4] Bekku, S. (1977), *Nihongo no Rizumu* (Rhythm in Japanese), Tokyo: Kodansha Gendai Shinsho.

[5] Poser, W.J. (1990), "Evidence for foot structure in Japanese," *Language* , vol. 66 (1), pp. 78-105.

[6] Teranishi, R. (1980), "Two-mora-cluster as a rhythm unit in spoken Japanese sentence or verse." Text of talk abstracted in *Journal of the Acoustical Society of America*, vol. 67, Supplement 1:S40.

[7] Minegishi, Y. (1958), *Uta-Awase no Kenkyu* (A Study of the Poetry Contests), Tokyo: Sanseido.

[8] Bownas, G. & A. Thwaite, (1964), *The Penguin Book of JapaneseVerse*, Penguin Books Ltd.

[9] Homma, Y. (1983), "The rhythm of *Tanka*, Short Japanese Poems," *Proceedings of the XIIIth International Congress of Linguists, Tokyo, Japan*, pp. 618-624.

COMPUTATIONAL MODELLING AND GENERATION OF PROSODIC STRUCTURE IN SWEDISH

Merle Horne and Marcus Filipsson
Department of Linguistics and Phonetics, University of Lund

ABSTRACT

A summary of the motivation for the various levels of structure assumed in a prosodic hierarchy for Swedish and the linguistic and discourse parameters that are needed for their recognition in texts are presented.

INTRODUCTION

Current speech synthesis systems, which lack detailed prosodic structure cannot generate many of the intonational patterns that one observes in natural speech. Prosodic phenomena associated with the boundaries of clause-internal word groups constitute one problem area. More specifically, the transitions that the current Swedish text-to-speech system generates between word accents do not always correspond to those one finds in naturally occurring speech. As the F0 curve in Figure 1 (corresponding to part of the sentence in (1)) shows, the end of the focussed expression *för närvarande* 'presently' coincides with a low F0 point (L#) in the speech of the radio commentator we are modelling. This L#, we claim below, corresponds to the end of the prosodic constituent which we will define as a [+focal] Prosodic Word. In Figure 2, it is observed that the corresponding synthetic F0 curve generated using the current rule system cannot reproduce this pattern since no low point after the focal high is predicted in clause-internal position. The F0 transitions are only triggered by the positions of the word accents which can be either focal (i.e. followed by a H⁻) or nonfocal (i.e. without an additional following H⁻). Thus, after the H*L (Accent 2) word accent on the syllable *-när-*, there is a rise throughout the remainder of the word (due to an associated focal H⁻) and the first syllable of the following word, *betecknas* 'is characterized', since the underlying accent pattern of an Accent 1 word like *betecknas* is HL*, with a H on the premainstress syllable *be-* and a L* on the syllable *-teck-* [1]. Thus, the L# at the end of *för närvarande* such as in Figure 1 cannot currently be automatically generated.

(1) För närvarande betecknas tendensen som mycket svag 'At present the trend is characterized as very weak'

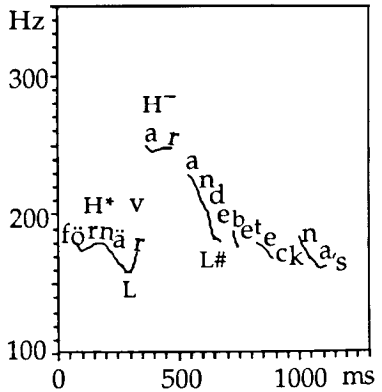


Figure 1. A partial F0 contour for the sentence in (1) uttered by a professional radio commentator.

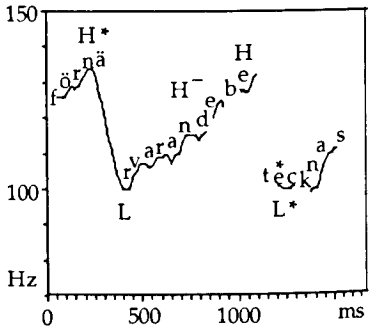


Figure 2. Synthesized F0 contour for the same sentence fragment as in Figure 1.

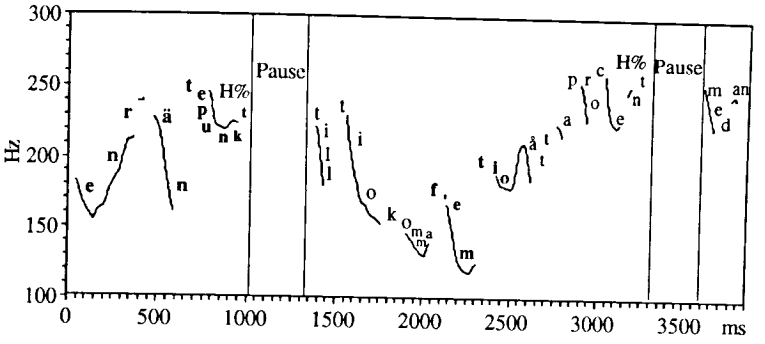


Figure 3. F0 contour for a fragment of the sentence in (2) (1 räntepunkt || till 10,58 procent || medan) with a clause-internal PP boundary after 'en räntepunkt'.

Another problem with current synthesis is that one has not been able to predict the location of clause-internal Prosodic Phrase boundaries, i.e. internal boundaries that are as strong as those which occur at the end of the majority of clauses/sentences. This is exemplified, for example, in Figure 3, which presents part of the F0 contour associated with the sentence in (2), where the internal boundary after *räntepunkt* 'interest point' has the same strength as that after *procent* 'percent'.

(2) ||*Tolvmånaders statsskuldväxlar hade gått tillbaka 1 räntepunkt || till 10,58 procent || medan sexmånadersväxlar gått upp 5 punkter till 10,50 procent*||
'Twelve-month state-debt bonds had gone back 1 point to 10.58 percent while six-month bonds had gone up 5 points to 10.50 percent.' (where || represents a Prosodic Phrase boundary)

In order to be able to recognize such internal Prosodic Phrase boundaries, one must have access to more lexico-grammatical information than is currently available in text-to-speech systems.

SWEDISH PROSODIC STRUCTURE

Prosodic Word (PW)

Three levels of prosodic structure are being assumed over the level of the syllable [2]. The smallest of these is the Prosodic Word (PW) which is defined as corresponding to a content word and any following function words up to the next content word within a given Prosodic Phrase (PPh). At the beginning of a PPh, the PW can also begin with one or more function words.

The PW is characterized by a word accent. It is also marked by a boundary tone which is realized by a final rise in the case where the content word is not focussed (i.e. contextually given) (H#) or a fall when the content word is focussed (L#). These boundary tones, we claim, play an important role in creating the transitions between consecutive PW's in a larger PPh. The unit does not necessarily correspond to a syntactic constituent; the grouping is, however, characteristic of well-planned speech. It is a rhythmic grouping with a left-headed character, where a content word can be grouped together prosodically with following function words in a manner analogous to the way the definite article and other morphological endings are attached and prosodically cliticized to the right of a lexical stem in Swedish (e.g. *bil+ar+na* 'car+pl.+the'). Thus, a PW can consist of a content word and a following preposition (e.g. *[köpt över]*, 'bought over') where the preposition is syntactically a member of a constituent that does not include the content word, as in *Lars har köpt över 100 skivor* 'Lars has bought over 100 records'.

Prosodic Phrase (PPh)

One or more PW's make up a PPh which is marked by a L% or H%

boundary tone accent, a following pause and a certain degree of Final Lengthening ([3]). It corresponds to both Pierrehumbert's 'Intonation Phrase' [4] and Lieberman's 'breath group' [5]. Factors which determine the location of PPh boundaries include the following:

a) clause/sentence boundary: A clause boundary corresponds in many cases to the end of a PPh. In an auditory analysis of a corpus of 36 radio broadcasts containing 724 clauses, where clauses also included elliptical clauses, 499 or 69% of these were characterized as ending in a boundary which was as strong or stronger than a PPh (404 were classified as PPh boundaries and 95 as Prosodic Utterance (PU) boundaries). Since we assume the Strict Layer Hypothesis in the hierarchy of prosodic constituents, this means that the end of a PU is also the end of a PPh; thus, 69% of the clauses ended in a boundary associated with a PPh at some level of analysis. Furthermore, 337 of these clauses corresponded to sentence boundaries. In the whole corpus, there were 362 sentences. This means that 93% of the sentence boundaries were assigned a prosodic boundary equal to a PPh on some level of analysis.

b) clause-internal PPh boundaries: Although PPh boundaries occur in the majority of cases in clause-final position, they may also occur optionally in clause-internal position. In our data of 724 clauses, we detected only 17 cases (2%). Although the number was extremely small, we decided, nevertheless to examine the lexico-syntactic structure of the data to determine whether one could make any generalizations concerning the environment for the insertion of these internal PPh boundaries. The following conclusions can be drawn from the investigation: in the domain-specific data-base dealing with the stock-market we investigated, 12 clause-internal PPh boundaries occurred before focussed complements (beginning with *till* 'to' or *sedan* 'since') to the verbs *gå upp* 'go up', *gå ner* 'go down', *falla* 'fall', *stiga* 'rise' IFF these complements were preceded by another focussed verb complement. Thus a PPh boundary (||) could occur before *till* in (4a) but not in (4b) where the first complement following the verb is not focussed (relevant focussed expressions are written in bold script):

(4) (a) ||*Fyra-åriga standardobligationen hade då fallit* ***4 punkter*** || *till en ränta på* ***10,27 procent***||
'||The four-year standard bond had fallen **4** points|| to an interest-rate of **10,27** procent||'

(b) ||*Tolvmånaders statsskuldväxlar hade också gått tillbaka 4 punkter till* ***10,84 procent***||
'||Twelve-month national-debt bills had also gone back 4 points to **10,84** procent||'

The remaining 5 cases of clause-internal PPh boundaries occurred between a relatively long subject (on the average of 15 syllables) and a focussed verb.

c) length: A PPh will consist of a more or less fixed number of syllables at a given rate of speech since PPh's (as we have defined them) correspond to what is often termed 'breath groups' [5]. In our material, where the speech rate is on the average of about 5 syllables/second, PPh's contained between 7 and 63 syllables, with the mean at 24 syllables (SD=10.3). Our data also indicate that a sentence-internal clause must be of a certain length in order for it to be associated with a PPh boundary; clauses containing less than 7 syllables (12 elliptical clauses) in our database were assigned a weaker, i.e. PW boundary. PW boundaries also replace PPh boundaries in a great many other cases. The 225 clause boundaries (31%) which were associated with a PW boundary instead of a PPh boundary all arose from the linking together of clauses within a discourse segment. Such linking never occurred over discourse segment boundaries. It was observed, futhermore, that clause-linking occurred only if the first clause contained less than 30 syllables and if the resulting PPh, after the linking of two or more clauses together, did not contain (on an average of) more than 40 syllables. The linking of clauses occurred most frequently at the beginning of discourse segments and practically never between the second last and final clauses of a discourse segment. Thus, it could be that the *non-linking* of clauses can be considered as a cue to segment finality (see also [6] for other cues).

Prosodic Utterance (PU)

One or more PPh's make up a PU, which is bounded by extended pauses [3]. These strong boundaries coincide with locations where a topic shift takes place (i.e. the end of a 'discourse segment' [7]). In our data, 95 of the 727 clauses (13%) ended in a boundary which was classified as a PU boundary. In the texts which were originally read on the radio, these correlate with the opening of a new paragraph immediately following the PU boundary (S. Haage-Palm, personal communication).

GENERATING PROSODIC STRUCTURE

In order to automatically generate prosodic structure, it is important to be able to recognize a number of different kinds of information [8]. First of all, the distinction between content words (e.g. nouns, adjectives) and function words (e.g. prepositions, conjunctions) is needed in order to define PW's. It is also crucial to be able to identify clause boundaries in a text since the clause is the basic domain over which PPh's are defined. Clause boundaries occur at certain punctuation marks, e.g. full stop, colon, semicolon, some commas (those not occurring in lists of words having the same word class), as well as before coordinate conjunctions and relative pronouns (e.g. *som* 'that') and after subordinate conjunctions (e.g. *att* 'that', *om* 'if').

In order to generate the clause-internal (optional) PPh boundaries, we have included a domain-specific module in our algorithm. This was due to the fact that the locations of clause-internal PPh boundaries seemed to be so domain-specific as regards their lexical specification. This is not the case with the module that identifies clause boundaries, which is domain-independent. Thus, the domain specific module inserts clause-internal PPh boundaries before the second focussed prepositional complement to the verbs *gå upp* 'go up', *gå ner* 'go down', *falla* 'fall 'and *stiga* 'rise'.

Moreover, in order to generate PW boundaries at the ends of clauses, i.e. in order to link two or more clauses together into a PPh, it is necessary to be able to calculate the number of syllables that a given clause consists of and the number of syllables that will result after its being linked to the following clause. This information is currently being built into the prosodic parser.

Finally, in order to generate PU boundaries, one must be able to recognize discourse segment boundaries. In the present algorithm [9] these are triggered by paragraph boundaries.

ACKNOWLEDGEMENT.

This research has been supported by a grant from the Swedish HSFR/NUTEK Language Technology Programme.

REFERENCES

[1] Bruce, G. and Granström, B. (1989), "Modelling Swedish intonation in a text-to-speech system", *STL-QPSR*, pp. 31-36.

[2] Horne, M. (1994), "Generating prosodic structure for synthesis of Swedish intonation", *Working Papers* (Dept. Ling., Univ. of Lund), pp. 43, 72-75.

[3] Horne, M., Strangert, E. and Heldner, M. (1995), "Prosodic boundary strength in Swedish: Final Lengthening and Silent Interval duration", *Proc. XIIIth ICPhS*, Stockholm.

[4] Pierrehumbert, J. (1980), *The phonetics and phonology of English intonation*. Ph.D. Diss., M.I.T.

[5] Lieberman, P. 1967. *Intonation, perception and language*, Cambridge, Mass.: MIT Press.

[6] Swerts, M. (1993), "On the prosodic prediction of discourse finality", Proc. ESCA Workshop on Prosody , *Working Papers* (Dept. Ling., U. of Lund) 41, pp. 96-99.

[7] Grosz, B. and Hirschberg, J. (1992), "Some intonational characteristics of discourse structure", *Proc. ICSLP 92*. Banff, pp. 429-432.

[8] Lindström, A., Horne, M., Svensson, T., Ljungqvist, M. and Filipsson, M. (1995), "Generating prosodic structure for restricted and "unrestricted" texts", *Proc. XIIIth ICPhS*. Stockholm.

[9] Horne, M. and Filipsson, M. (1994), "Generating prosodic structure for Swedish text-to-speech", *Proc.* ICSLP 94, Yokohama, pp. 711-714.

THE ROLE OF LEXICAL STRESS IN THE RECOGNITION OF SPOKEN WORDS: PRELEXICAL OR POSTLEXICAL?

Willy Jongenburger and Vincent J. van Heuven
Dept. Linguistics/Phonetics Lab., Leiden University/
Holland Institute of Generative Linguistics, The Netherlands

ABSTRACT

In this study we investigate to what extent lexical stress information is used to narrow down the cohort of potential word candidates. Our gating data on Dutch minimal stress pairs showed that lexical stress information is not used in the activation phase of the word recognition process, but does contribute to the prelexical selection stage.

INTRODUCTION

In stress-accent languages such as English, German and Dutch the position of the stressed syllable varies from one word to the next. Information on the position of the stressed syllable might contribute to the human word recognition process. Very few studies investigated to what extent lexical stress narrows down the cohort of potential word candidates and so far, the experimental data present a confusing picture. There is evidence that the context preceding an accented monosyllabic word contains prosodic cues about the stress pattern of this word: the melodic and rhythmic organisation of the preceding context tells the listener when to expect an accented syllable [1,2].

Dutch listeners performing a gating task with isolated words under optimal conditions only need the first syllable of the target word to know whether this syllable is (lexically) stressed or not. In LP filtered speech (750Hz, -48dB/oct), however, there was a strong bias for initially stressed responses, both for initially stressed targets and for initially unstressed targets [3]. Gating and shadowing experiments showed that stressed versus unstressed realisations of otherwise identical word-initial full syllables effectively narrowed down different cohorts of recognition candidates [4]. A control experiment [5] justified the conclusion that lexical stress realised on the target words is used in the early word recognition process, and that the relevant cues are provided by the first syllable of the target string, rather than by the prosody of the preceding context. So, prosodic information, notably the difference between stressed and unstressed but segmentally identical word onsets, is used in the word recognition process.

These findings contradict claims for English that the effects of stress are located in the postlexical phases of the word recognition process only [6]. During the prelexical activation and selection phase minimal stress pairs in English proved to be functional homophones in an on-line cross-modal priming experiment.

Since gating provides information about the cohort of word candidates at different stages in the word recognition process, we ran gating experiments with Dutch minimal stress pairs in context, using not only high-quality speech but also segmentally degraded (LP filtered) speech. In segmentally degraded speech, the relative contribution to word recognition of segmental and prosodic information changes. Typically slowly varying prosodic information is more resistant to distortion of the speech signal than relatively fast varying segmental information. In LP filtered speech the time-span of the prelexical phase is increased, so that stress information gets a better chance of contributing to the prelexical recognition phases.

GATING STUDY

Our gating study included two conditions: hifi speech and LP filtered speech. In order to obtain speech of poor quality that is still sufficiently intelligible, a pilot study was carried out to establish the appropriate cut-off frequency for LP filtering (-48dB/oct). Individual cut-off frequencies were established for each target word, so as to guarantee that the two members of a minimal stress pair are equally (un)intelligible. Individual cut-off frequencies varied from 1250 Hz up to 3000 Hz.

METHOD

A male speaker of standard Dutch recorded seven Dutch minimal stress pairs. The two members of each pair were embedded in a non-biased (semantically neutral) sentence as in:

Ze dacht dat haar vriend CAnon/kaNON opzocht.
she thought that her friend canon / gun looked up

With the aid of a waveform editor these utterances were cut into fragments of increasing length, under visual and auditory control. For both quality and stress conditions the same truncation points were chosen. The first gate consisted of the preceding context plus the initial consonants and the first vowel onset of the target word. Each next fragment contained one diphone more, i.e. the second fragment included the initial consonant(s), vowel and onset of the next consonant, until the whole word and even the beginning of the next word had been gated. For each speech condition two stimulus series were prepared with the stress pattern of the target words counterbalanced; each series contained one member of each minimal stress pair.

Forty subjects participated; each stimulus series was presented on-line over headphones to ten subjects. Subjects were instructed to write down and say aloud after each fragment, the word they thought was being presented. They also had to indicate on a 10-point scale how confident they were as to the correctness of their response.

RESULTS AND DISCUSSION

In order to investigate to what extent lexical stress helps the listener to narrow down the cohort of potential word candidates, written responses were analysed. In cases where the orthographic responses did not allow us to unambiguously establish the stress pattern of the responses, the audio recordings were analysed instead. Monosyllabic content words were considered initially stressed, monosyllabic function words as initially unstressed. An Anova on the isolation point data shows only a large main effect of speech quality condition $F(1,214)= 15.1$, $p<0.001$. Subjects need more acoustic information to isolate the target in LP filtered speech then in high-quality speech: 5.0 versus 4.0 gates. This means that the time span of the prelexical phase in LP filtered speech is indeed increased. Figure 1 presents the cumulative distributions of percent correct word responses and of initially stressed error responses as a fuction of gate length, broken down by the stress pattern of the target, for high-quality and LP filtered speech.

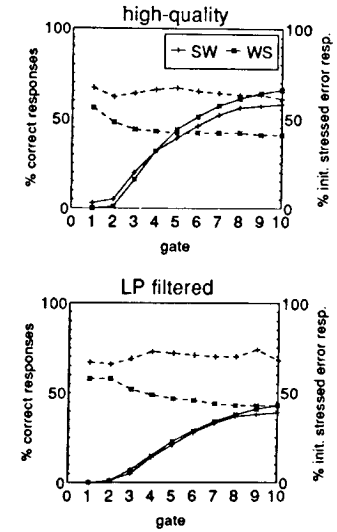


Figure 1. Percent correct word responses (solid lines) for high quality (upper panel) and LP filtered (bottom panel) speech and initially stressed error responses (dotted lines), as a function of gate length, broken down by the stress pattern of the target (SW versus WS).

For high-quality speech the percentage of initially stressed error responses obtained for initially stressed (SW) targets is

clearly higher than for initially unstressed (WS) targets at all gates. The difference in the proportion of initially stressed error responses between SW and WS targets is statistically significant at gates 3, 4 and 5 ($\chi^2 \geq 5.1$, $p \leq 0.02$. At earlier gates the differentiation is insignificant, at later gates the number of cases is too small to run statistical tests.

As to LP filtered speech it appears again that the proportion of initially stressed error responses is considerably larger for SW targets than for WS-targets. From the first gate onwards, stressed and unstressed word beginnings lead to different distributions of error responses. The differentation assumes statistical significance (χ^2 between 5.1 and 19.2 for gates 3 through 8, df=1, $p<.05$) from gate 3 onwards, but, crucially, it is again insignificant during the first two gates, i.e. during the presentation of the initial syllable.

The observation that lexical stress is not heard during the first two gates suggests that prosodic information is not used during the activation phase. From the third gates onwards, however, listeners might use lexical stress in the selection phase. One may ask why the present results show only moderate differentiation of stressed and unstressed word beginnings, whilst much stronger differentiation was reported in the literature [2,3]. Several answers spring to mind. First, in the present experiments low-predictability words were embedded in uninformative contexts, whereas more easily available words were used in the earlier experiments, where they occurred either in isolation or in a slot in a carrier phrase. Second, in the earlier experiment with carrier sentences, listeners were provided with a typed version of the sentence up to the critical word, whereas our subjects had no information about the position of the target's onset: i.e. listeners did not know beforehand that the fragment ended with a word onset.

We will now try to show that lexical stress information is actually used in the selection phase of the word recognition process. If at the gate preceding the isolation of the target, i.e. one gate before the subject produces the target word without subsequently changing his mind, the proportion of rhythmically correct error responses is considerably larger than at the same gate for all cases where isolation of the target does not follow, this would be an indication that prosodic information does in fact constrain the cohort of word candidates. Figure 3 presents the rhythmically correct error responses at the gate immediately preceding the isolation point for all cases where the listener did in fact reach an isolation point (+iso) as well as the corresponding rhythmically correct error responses at the same gate number when no subsequent isolation of a target followed (-iso).

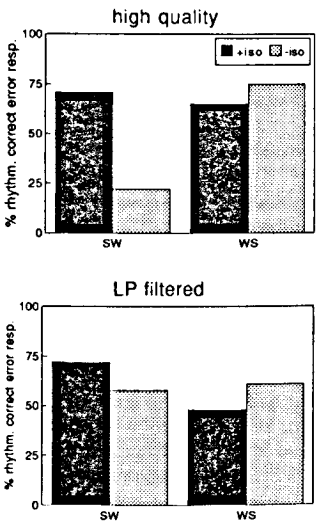


Figure 2. Percent rhythmically correct error responses for high-quality (upper panel) and LP filtered speech (lower panel) at the gate preceding isolation (+iso) and at the same gate position without isolation following (-iso).

The proportion of correctly stressed error responses in high-quality speech at the gate preceding isolation (+iso) for SW targets is considerably larger than at the same gate when no isolation of the target follows (-iso); $\chi^2=7.7$, $p=0.006$. This means that listeners use prosodic information in the early phases of word recognition. The proportion of rhythmically correct error responses evoked by high-quality WS-targets does not differentiate between +iso and -iso.

Although in LP filtered speech the proportion of correctly stressed error responses evoked by SW targets is larger for +iso than for -iso, this difference is insignificant.

In LP filtered speech the proportion of correctly stressed error responses evoked by WS targets in the -iso condition is statistically the same as in the +iso condition. This means that although listeners hear an unstressed initial syllable, they still are willing to reconsider a stressed onset syllable. These observations are in accordance with earlier findings [7] that unstressed syllables are not generally used by Dutch listeners to eliminate recognition candidates that begin with a stressed syllable, but that hearing stressed onset syllables effectively block access to that part of the mental lexicon containing initially unstressed words.

CONCLUSIONS

We will now recapitulate the main points of this paper. 1) Listeners proved unable in a gating task to differentiate between stressed versus unstressed beginnings of minimal stress pairs as long as no larger onset portion of the target was made audible than the first syllable. 2) Differentiation increased *after* the first syllable. Moreover, the subjects' word recognition was shown to be facilitated when the rhythmic pattern was correctly perceived before the isolation point was reached. We suggest, on the basis of these findings that lexical stress information is not used in the activation phase of the word recognition process, but still contributes to the prelexical selection stage.

Since the validity of the gating paradigm as a simulation of the on-line recognition process is subject to discussion, it is difficult to interpret whether these data falsify Cutler's claim that stress information does not play any role at all in lexical acces. Therefore we are currently running (on-line) cross-modal priming experiments [8] with the same experimental material used in this gating study.

ACKNOWLEDGEMENT

We acknowledge the financial support given by the Netherlands Organisation for Research (NWO) through the Foundation for Speech, Language and Logic, under project # 300-173-023.

REFERENCES

[1] Cutler, A. (1976), Phoneme-monitoring reaction time as a function of preceding intomation contour, *Perception & Psychophysics* 20, pp. 55-60.

[2] Cutler, A. & Darwin, C.J. (1981), Phoneme monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency. Perception & Psychophysics 29, pp217-224.

[3] Heuven, V.J. van (1984), Verslagen van de Nederlandse Vereniging voor Fonetische Wetenschappen, No. 159-162, pp 22-38.

[4] Heuven, V.J. van, (1988), Effects of stress and accent on the human recognition of word fragments in spoken context: gating and shadowing, *Proceedings of the 7th Fase/Speech-88 Symposium*, Edinburgh, pp 811-818.

[5] Dalen, J. van & Noorloos, J. van (1989), De rol van klemtoonmarkering bij de lexicale access, unpublished paper, Leyden University.

[6] Cutler, A. (1986). Forbear is a homophone: Lexical prosody does not constrain lexical access, *Language and Speech*, 29, pp. 201-220.

[7] Heuven, V.J. van, (1985). Perception of stress pattern and word recognition: Recognition of Dutch words with incorrect stress position, *Journal of the Acoustical Society of America*, 78, s21.

[8] Jongenburger, W. & Heuven, V.J. van, (1995), The role of linguistic stress in the time course of word recognition in stress-accent languages, *Proceedings of Eurospeech '95* (to appear).

The Phonetic Basis of Phonological Foot: Evidence from Japanese

Haruo Kubozono
Osaka University of Foreign Studies, Japan
(g01161@sinet.ad.jp)
University of California, Santa Cruz, U.S.A.

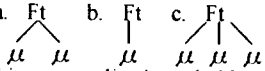
ABSTRACT

It is widely recognized that bimoraic foot represents the most canonical foot form in the phonological theory of stress and rhythm. This paper explores the phonetic basis of bimoraic units in phonological descriptions by analyzing some phonetic, phonological and lexical phenomena in Japanese from various viewpoints including historical phonology, comparative phonology and language acquisition.

INTRODUCTION

It is generally agreed that bimoraic foot represents the most unmarked foot structure. In the metrical typology of linguistic rhythm [1], for example, quantity-sensitive languages can take one of the two metrical principles, iambic or trochaic, for both of which bimoraic feet (1a) represents a better configuration than monomoraic (1b) or trimoraic feet (1c).

(1) a. Ft b. Ft c. Ft



This generalization holds true of Japanese, where the notion 'bimoraic foot' allows us to generalize a wide range of phenomena from morphological to pitch-related processes [2]. Although there is some evidence for monomoraic feet in (1b) in this quantity-sensitive system [3], bimoraic feet are no doubt much more unmarked than monomoraic feet.

While the notion 'foot' is known to be indispensable for morphological and phonological descriptions in general, it is by no means clear why such a phonological unit exists, specifically why 'bimoraicity' rather than monomoraicity or trimoraicity represents the most canonical foot form in language in general. Since the average duration of the Japanese mora is about 150 msec in normal speech [4], the average duration of bimoraic feet is about 300 msec. The question we should ask, then, is whether bimoraic foot phenomena in Japanese and other languages have to do with this phonetic duration. In this paper I will explore this possibility through analysis of the phonetic structure of bimoraic foot phenomena in Japanese from a historical, psycholinguistic and cross-linguistic viewpoints.

The key to the question of bimoraicity lies in the fact that many, if not all, of the bimoraic foot phenomena in Japanese serve to create heavy (i.e. bimoraic) syllables as opposed to light (monomoraic) and superheavy (trimoraic) syllables. Since the notions 'bimoraic foot' and 'heavy syllable' both involve integrating two moras into one phonological unit, the question of phonological bimoraicity may be linked to the question of why bimoraic syllables are preferred to monomoraic and trimoraic syllables.

DURATION OF RHYTHMIC FEET

The first fact to note regarding the phonetic duration of phonological foot is that stress feet in English and other languages take a similar phonetic duration as the bimoraic feet in Japanese. According to the experimental work by Dauer [5], inter-stress intervals in so-called stress-timed and syllable-timed languages alike fall within the range of 300-600 msec. This tendency may be interpreted as an effect of the 'neural clock' [6], a clock that strikes about two times per second in human mind. If this interpretation is correct, it follows that the Japanese mora is too short to form a rhythmic unit by itself which, in turn, suggests that Japanese has chosen to combine two moras into one unit in order to create a eurhythmic prosodic structure out of the otherwise monotonous sequences of moras. One question that remains is why Japanese does not choose to form trimoraic feet, or a prosodic unit of about 450 msec in phonetic terms: this would be closer to the average duration of the rhythmic feet in other languages and, hence, to the duration dictated by the 'neural clock'. We return to this question at the end of the paper.

CHILD PHONOLOGY

A second piece of evidence for the phonetic interpretation of phonological foot comes from a phonological analysis of children's speech in Japanese. Children's language as well as the language used by adults when addressing to young children in Japanese is crucially different from adults' language per se in containing an extremely high proportion of heavy syllables. This tendency shows up in three independent ways. First, children's vocabulary typically contains words rich in heavy syllables. According to my previous analysis [7], heavy syllables account for fifty percent (or more) of all the syllables occurring in the words spoken by two- to three-year-old children. In fact, words used at this early stage of phonological development typically take one of the two syllable structures in (2): /.I denotes a syllable boundary.

(2) a. Reduplication of a heavy syllable
hai.hai "crawling" pon.pon "belly"
tin.tin "penis"
b. A heavy syllable + a light syllable
man.ma "food" kuk.ku "shoes"
an.yo "foot, leg, walking"

This fact contrasts sharply with the fact about adult speech, where bimoraic syllables account for less than ten percent of all the syllables occurring in natural speech (the rest being light, i.e. monomoraic, syllables).

Secondly, young children show a strong tendency to choose words containing heavy syllables rather than light syllables. According to my own observation, children prefer the words in (3a) to their synonymous counterparts in (3b):

(3) a. is.sai "one year old"
ni.sai "two years old"
san.sai "three years old"
b. hi.to.tu "one year old"
hu.ta.tu "two years old"
mit.tu "three years old"

The preference of bimoraic syllables in (2) and (3) suggests that bimoraic syllables are somehow easier for children to pronounce. This might be taken as implying that it is the phonological structure of heavy syllables and not their phonetic duration that is relevant. However, this interpretation can be refuted by the fact in (4), where it is shown that one- to two-year-old children often lengthen light syllables, showing a free variation between short and long vowels. This suggests that young children at this stage of phonological development adjust the phonetic size of syllables in an attempt to attain the bimoraic duration at the phonetic level.

(4) /ni.sai/ "two years old"
→ [nisai] ~ [ni:sai]
/o.too.san/ "father"
→ [oto:san] ~ [o:to:san]

Interestingly, the tendency toward heavy syllables in (2)-(4) is not a language-specific phenomenon. Allen & Hawkins [8] note a similar tendency with respect to English-speaking children: "children's earliest phonologically patterned untteracnes typically have only heavy syllables.... The development of light syllables therefore represents...an important step in the child's development toward adult phonological rhythm" (p.274).

ADULT PHONOLOGY

It is important to emphasize here that the tendency toward establishing syllables of a bimoraic size is observed in adult language too. In fact, many of the phenomena cited for the phonological notion of bimoraic foot in the literature involve a phonetic lengthening by which bimoraicity is achieved at the phonetic level. For example, many monomoraic content words are lengthened into heavy syllables in their citation form, as illustrated in (5) ([9]). These lengthening processes seem to be phonetic rather than phonological in nature since speakers are generally unaware of the vowel lengthening involved and do not reflect it in writing (note that (5c) occurs only in some dialects of Japanese [4]). Here, again, 'bimoraic foot' phenomena attempt to attain bimoraicity at the phonetic level.

(5) a. days of the week
/ka moku do/ → [ka: moku do:]
"Tuesday, Thursday and Saturday"
b. numbers
/ni go roku/ "256" → [ni: go: roku]
c. nouns
/me/ "eye" → [me:]
/te/ "hand" → [te:]

HISTORICAL EVIDENCE

In addition to the synchronic evidence from modern Japanese, some historical evidence reinforces the idea that phonological units of a bimoraic size are motivated by the temporal structure of speech. Although it is difficult to trace the history of bimoraic feet in Japanese, it is nevertheless possible to trace the history of bimoraic syllables in the

language.

The tendency to create bimoraic syllables is not recent in the history of Japanese. Although Japanese in Nara Period (eighth century) supposedly had only one syllable structure, i.e. CV, and no contrastive vowel or consonant length, it subsequently established heavy syllables (CVC and CVV) as a legitimate syllable structure. Apart from the influence of Sino-Japanese vocabulary which was rich in this second type of syllable structure, the most noticeable internal change which contributed to the development of bimoraic syllables in Japanese is a series of sound change known as *onbin*, which started in early Heian Period (ninth century). This sound change had the effect of converting a sequence of light syllables into a heavy syllable by way of deletion of a consonant or a vowel:

(6) a. tu.ki.ta.ti → tui.ta.ti
"the first day of the month"
b. yo.mi.te → yon.de "to read"

Given this syllable-based account of *onbin*, one may naturally ask why heavy syllables became a legitimate syllable structure suddenly at this stage of the history of Japanese or, stated conversely, why only CV syllables were tolerated in the pre-*onbin* period. The key to this question lies in the distinction between phonological quantity and phonetic duration of the syllable.

The CV syllable in Old Japanese is supposed to have been phonetically much longer than the CV syllable in modern Japanese for several independent reasons. First of all, Old Japanese was closer to a tone language like modern Chinese than a pitch accent language like modern Japanese in terms of the number of distinctive pitch contrasts [10]. Since syllables tend to be longer in duration in a tonal system than in a pitch-accent system — e.g. the average duration of syllables in modern Chinese is reported to be about 450 msec [11], which is three times as long as the average CV syllable in modern Japanese — it can be assumed that the CV syllable in Old Japanese was much longer than the CV syllable in modern Japanese. Secondly, there is historical and synchronic evidence that monosyllabic content words were much longer in Old Japanese than they are in modern Japanese. This is evidenced by historic documents in which monosyllables are transcribed as possessing a bimoraic duration and also by the synchronic fact that monosyllables in the more classical and conservative dialect of Kyoto and Osaka are nearly twice as long as their counterparts in Tokyo Japanese [4].

This leads us to assume that CV syllables in Old Japanese were phonetically more equivalent to heavy syllables (CVV and CVC) than to light syllables (CV) in modern Japanese and, hence, that they were phonetically as well-formed as heavy syllables in quantity-sensitive languages. Seen in this light, *onbin* and other phonological processes responsible for the creation of bimoraic units (heavy syllables and bimoraic feet) can be analyzed as being triggered by the shortening of phonetically long CV syllables. In terms of rhythmic regulation of speech, this means that both the tendency to create heavy syllables and to group two moras into one rhythmic foot, whether at the phonetic or phonological level, can be attributed to a force that imposes bimoraicity on phonological material at the phonetic level of speech.

LOANWORD PHONOLOGY

Before concluding this paper, let us consider the question of why bimoraic feet are generally more favored than trimoraic feet. This question is difficult to answer, but it may be tackled from the viewpoint of the canonical quantity of the syllable. In addition to the phonetic processes described in (4)-(5) above, Japanese exhibits many phonological processes by which bimoraic syllables are established. In loanword phonology, for example, obstruents following a short stressed vowel in the source language are generally geminated in Japanese and, together with vowel epenthesis, produce a sequence of a heavy syllable followed by a light syllable:

(7) Consonant gemination
cup → /kap.pu/
back → /bak.ku/
push → /pus.sju/

However, this phonological adjustment is blocked if the preceding vowel is a long vowel or diphthong, i.e. if consonant gemination would produce a trimoraic syllable rather than a bimoraic syllable.

(8) Antigemination
carp → /kaa.pu/, */kaap.pu/
baiku → /bai.ku/, */baik.ku/

The stressed syllables in (7) result in a heavy syllable by undergoing gemination while those in (8) attain the same syllable quantity by NOT undergoing it. Thus the two phenomena in (7) and (8) have the same target, i.e. phonological creation of heavy syllables.

'Pre-nasal shortening' [12] produces the same effect by shortening long vowels and diphthongs followed by a nasal /n/ in the process of borrowing. This shortening too has the effect of yielding bimoraic syllables in contexts where trimoraic syllables would otherwise be created.

(9) ground → /gu.ran.do/, */gu.raun.do/
angel → /en.zje.ru/, */ein.zje.ru/

The processes in (8) and (9) are particularly interesting in suggesting that trimoraic syllables are as well as monomoraic syllables are marked in Japanese. Again, this is not a language-specific phenomenon but is observed in a variety of languages: see, e.g. [13]. This hints that three moras cannot be easily accommodated into one unit, which may be linked to the fact that trimoraic feet in (1c) are disfavored in the organization of phonological rhythm.

CONCLUDING REMARKS

In this paper I discussed the phonetic nature of phonological foot by presenting, among others, the following two lines of evidence. First, analysis of young children's speech shows two marked tendencies: (i) dominance of phonologially heavy syllables (CVV and CVC) over light syllables (CV), and (ii) lengthening of monomoraic syllables, consequently making syllables of this type equivalent to heavy syllables at the phonetic output of speech. Second, cross-linguistic and historical considerations reveal that CV syllables in Old Japanese were phonetically much longer than light syllables in modern Japanese, and that heavy syllables were established in the language immediately after CV syllables were phonetically shortened due probably to some independent prosodic factors. All these observations can be generalized if it is hypothesized that 'bimoraicity' can be achieved by phonetic means (phonetic lengthening of light syllables) as well as by phonological means (preference of heavy syllables), which, in turn, seems to suggest that 'bimoraicity' embodies a phonetic requirement on the duration of the syllable rather than any formal phonological requirement on the size of the syllable.

The notion 'bimoraic foot' naturally follows from this interpretation in such a way that establishing the bimoraic foot as a phonological unit is another way of establishing a domain where bimoraic duration (about 300 msec) is achieved at the phonetic level of speech. This phonetic notion of bimoraicity can be linked in a natural manner to the phonetic 'stress foot' in English and other languages, which is known to take a similar duration of time.

ACKNOWLEDGEMENT

This work was supported by a 1994-5 U.S. government grant (Fulbright fellowship) for junior researchers.

REFERENCES

[1] Hayes, B. (1995) *Metrical Stress Theory: Principles and Case Studies*. Chicago: The University Chicago Press.
[2] Poser, W. (1990) "Evidence for foot structure in Japanese", *Language* 66: 78-105.
[3] Kubozono, H. (1995) "Degenerate feet in Japanese", ms. UC Santa Cruz.
[4] Sugito, M. (1989) "Onsetsu ka haku ka (syllable or mora)", in M. Sugito (ed.) *Nihongo no Onsei Onin*. Tokyo: Meiji-shoin.
[5] Dauer, R.M. (1983) "Stress-timing and syllable-timing reanalyzed", *Journal of Phonetics* 11: 51-62.
[6] Allen, G.D. (1973) "Speech rhythm: Its relation to performance universals and articulatory timing", *Journal of Phonetics* 3: 78-86.
[7] Kubozono, H. (1993), The syllable in Japanese, m.s. Osaka University of Foreign Studies.
[8] Allen, G.D. and S. Hawkins (1978) "The development of phonological rhythm", in A. Bell & J.B. Hooper (eds.) *Syllables and Segments*. Amsterdam: North-Holland, 173-185.
[9] Itô, J. (1990) "Prosodic minimality in Japanese", *CLS 26-II: Papers from the Parasession on the Syllable in Phonetics and Phonology*, 213-239.
[10] Komatsu, H. (1981) *Nihongo no Onin (Japanese Phonology)*. Tokyo: Chuo-koronsha.
[11] Mochizuki, Y. (1983) *Chugokugo to Nihongo (Chinese and Japanese)*. Koseikan.
[12] Lovins, J.B. (1975) *Loanwords and the Phonological Structure of Japanese*. Indiana University Linguistics Club.
[13] Myers, S. (1987) "Vowel shortening in English", *NLLT* 5: 485-518.

PHONETIC CUES AT SENTENCE BOUNDARIES IN QUEBEC FRENCH: A SIDE EFFECT OF PENULTIMATE SYLLABLE DURATIONS?

M. Ouellet and J. Lavoie
CIRAL, Laval University, Québec, Canada

ABSTRACT

This study is concerned with the characteristics of long penultimate syllables in Quebec French spontaneous speech and the way this typical duration seems to model the realization and interpretation of phonetic cues at prosodic boundaries, that is in sentence-final and phrase-final position.

INTRODUCTION

It is an established fact that « pretonic lenghtening » plays a part in the rhythmic patterning of Canadian French. [1,3,11]. Initially associated to an extended use of the *accent d'insistance* by Fouché for French from France, then by Boudreault for Canadian French, the phenomenon of pretonic lenghtening is now considered as ensuing from a more complex dynamic, involving segmental durations, syllable structure and morpheme boudaries [1,2,4,5,11].

One of the most important source of penultimate lengthening arises from vowel phonology. The vowel system of Canadian French has maintened eight intrinsically long vowels: /ɑ o ø ɜ ɜ̃ ɛ̃ ɑ̃ œ̃/. Those vowels can remain long in pretonic position, in a closed as well as in an opened syllable. Besides, lenghtening rules, which systematically apply under stress, become optional otherwise [7]. Intrinsically short vowels which are obligatorily lenghtened by /ʒ ʁ v z/ in closed stressed syllables sometimes stay lenghtened in non-final position if they belong to a morpheme (*pire* [pi:ʁ], *empirer* [ɑ̃pi:ʁe]). Lengthening rules are, in fact, sensitive to the presence of morphological boundaries [4,7,10].

In this first part of a larger study dealing with interactions between high and low level of constraints in spontaneous Quebec French, we tried to define the nature of the phonological composition of long penultimate syllables. A second aim was to evaluate the repercussion of those long penultimate syllables on stress patterning of Quebec French, that is to say a possible stress shifting from final to penultimate syllable in spontaneous speech.

METHODS

Among the available data on the structure and the effect of long penultimate syllables, only few come from the acoustical analysis of spontaneous speech [8]. That is the reason why we gathered a corpus of 108 sentence-final and phrase-final excerpts, extracted from 8 sociolinguistic interviews (ref.[6]) in which long penultimate syllables were perceived by three trained listeners. We kept only sequences for which the three judges agreed on the presence of a long syllable. Sentences were digitized (20kHz) with CSL program. Duration and frequency measurements were performed on strings of six syllables, strarting from the end of the sentence or phrase. Finally, the same listeners were asked to determine stress placement.

In order to describe the composition of long penultimate syllables, we examined factors such as intrinsic vowel duration, nature of neighboring consonants, syllable structure (closed or opened) and presence of a morpheme boundary. Comparisons between penultimate and final syllables were based on syllable durations, vowel durations, relative duration of nucleus (syllable portion hold by the vowel expressed in percentages), number of phoneme per syllable, rising or falling pitch and intonational scope (flat intonation: less than 1.5 semitone; minor scope: between 1.5 and 3 semitones; and major scope: more than 3 semitones) [9]. Sentence-final or phrase-final position of the strings and stress placement were also examined.

RESULTS

Long Penultimate Syllable Composition

More than 65% of long penultimate syllables contained an intrinsically long vowel such as /ɑ o ø ɜ ɜ̃ ɛ̃ ɑ̃ œ̃/. An interesting aspect of those results is brought by the relatively high frequency of short vowels in long penultimate syllables. The short vowel lenghtening rule implies /ʒ ʁ v z/ following the vowel in a closed syllable or inside a morpheme. This rule actually explained only 6 cases out of 43 in the sample. One might consider that, somehow, short vowels simply behave like long vowels in preserving their own durational quality in penultimate syllables. Moreover, lax allophones of high vowels [I Y U] never appeared as nucleus of long penultimate syllable. Laxed allophones of /i y u/ were produced in closed syllables, preceeding all consonants but /ʒ ʁ v z/. [I Y U] can appear in non-final opened syllable according to an optional laxing harmony rule [7, 11].

Table 1. Distribution of vowels in penultimate long syllables

nasal	35,2%
long oral	30,6%
short oral	25,0%
tense [i y u]	9,3%
lax [I Y U]	0%

All long penultimate syllables but two had at least one consonant as onset (that consonant may come from preceeding words or syllables). Out of that number, less than 20% were branching onsets. Even if the preceeding consonant has only a weak influence on vowel durations in Quebec French, [4] we compiled a list of those, which appeared before the syllable nucleus.

Table 2. Preceeding consonants

/ l / or / r /	28%
voiceless stop	25%
nasal	20%
voiceless fricative	9%
voiced stop	7,5%
voiced fricative	7,5%
approximant /j ɥ w/	3%

Table 3. Following consonants

voiceless stop	32%
voiceless fricative	26%
voiced fricative	15%
/ l / or / r /	14%
voiced stop	6%
nasal	4%
approximant /j ɥ w/	3%

According to the compilation of Table 3, only 29% of all vowels were followed by consonants which may promote vowel lenghtening . However, the effect of the following consonant had to be evaluated according to the place of syllable boundary. Our data provided 90 opened syllables and 18 closed ones. Out of those 18 syllables, only two had a branching coda. The huge number of opened syllables indicates that structure ensuing from syllabification rules in French does not explain the production of long penultimate syllables: first, the lengthening effect of following consonants is mitigated if they belong to another syllable [4]; second, the long penultimate syllable should not be

conceived has including systematically a coda.

Nevertheless, a morpheme boundary was encountered in 57 syllables (including those 18 closed syllables already mentioned). Then, considering the morphological conditioning in syllabification, 53% of long penultimate syllables might be closed ones, structurally or morphologicaly.

Penultimate and final syllables

Difference in syllable durations cannot explain the perception of long penultimates since only half of them were, in fact, longer than final ones. On the basis of the number of phonemes per syllable, 19 penultimates were longer than final syllables. Again, this factor cannot explain the longer penultimates.

On the other hand, vowel absolute and relative durations might be a perceptual clue: 74% of penultimate vowels were longer than final vowels. If we consider the relative duration of vowels, this proportion goes up to 78%.

Intonational Patterning.

A rising movement of frequency seemed to characterize sentences and phrases containing a long penultimate syllable since that pattern was found for 79% of the sample. Intonational scopes in semitones were distributed as follows: less than 1.5 semitone 32%, minor scope 26%, major scope 42%. We found a falling frequency pattern on 23 strings. Intonational scopes were distributed as follows: less than 1.5 semitone 56%, minor scope 9% and major scope 35%. Now, regardless of rising or falling pattern, major scopes were the most frequent (41%), followed by flat intonation (37%), and by minor scopes (22%).

Stress Placement

Even though the entire corpus was made of long penultimates syllables, stress was perceived on final syllables most of the time (78%). With the intent to identify the factors apt to provoke the placement of stress on final syllables, we performed a binomial variable rule analysis (Varbrul), including all independent variables but those relating to the identity of phonemes. The only significant factor was intonation (.000). Regardless of the intonational scope (non significant .592), rising intonation promoted stress perception on the final syllable (rule weight: .66). On the other hand, falling intonation was closely related to stress placement on penultimate syllables (significance: .000; rule weight: .92).

The sample gathered included 73 sentence-final strings and 35 phrase-final strings. Those proportions should be strictly attributed to the sampling since we looked first to gap sentence-final strings. Besides, the factors related to position of the string remained non-significant in binomial analysis.

DISCUSSION

The description of long penultimate syllables composition lead us to the conclusion that the most reliable and consistent indication seems to be absolute and relative vowel durations. As second factor, the influence of morphology might be considered as playing a part in the perception of syllable weight. These results raise further questions about the emergence of short vowels as nucleus of long penultimate syllables. Assumption of an expanding use of long vowel lengthening rule in non-final syllables is still to be investigated.

Cues may arise from the analysis of following consonant durations (Table 3). Penultimate long vowels are regularly followed by voiceless consonants which may play a part in penultimate perceived lenght, since they are usually longer than other consonants. If that perceived lenght results from an evaluation of the distance between penultimate and final nucleus PCenters, the consonant duration might provoke an interpretation of penultimate as being a closed syllable.

The second part of this study referred to a possible stress shifting from final to penultimate syllable. Surprinsingly, that shifting happens only when the intonation falls. At the opposite, stress is still perceived on final syllable with a rising pattern. This phenomenon already noticed in Quebec French seems to originate from a strategy developed to solve the problem brought by having two adjacents temporal cues: lenght in penultimate and final syllables [10]. Then, if duration cannot provide cues for stress placement, intonation does. Should we talk now of pitch shifting instead of stress shifting?

The resuts we obtained show clearly that phonological duration may interfere with prosodic constraints. The rising pattern could not be associated with social factors since there is no consistency between the social characteristics of the speakers and the distribution of that pattern, nor with the production of long penultimate syllables. It is possible that this inconsistency be attributed to an insufficient number of speakers. Actually, it is not possible to rule on the expanding nor on the recessive nature of penultimate lengthening phenomenon in Canadian French.

ACKNOLEDGEMENTS

Research supported by the Social Sciences and Humanities Research Council of Canada under grant no 410-94-1371.

We would like to thank Benoît Tardif for his tremendous help in performing part of the acoustic analysis.

REFERENCES

[1] Boudreault, M. (1968), *Rythme et mélodie de la phrase parlée en France et au Québec*, Québec, Presses de l'Université Laval.

[2] Fouché, P. (1956), *Traité de prononciation française*, Paris.

[3] Gendron, J.-D. (1966), *Tendances phonétiques du français parlé au Canada*, Québec, Presses de l'Université Laval.

[4] Ouellet, M. (1992), *Systématique des durées segmentales dans les syllabes en français de France et du Québec*, PhD Dissertation, University of Montreal.

[5] Ouellet, M. (1994), « Faits systématiques de durée en français québécois, rythme et accentuation. », *Actes des XXèmes Journées d'étude sur la parole*, pp. 367-372.

[6] Paradis, Cl. (1985), *An acoustic study of variation and change in the vowel system of Chicoutimi and Jonquiere (Quebec)*, PhD Dissertation, University of Pennsylvania.

[7] Paradis, Cl. (1992), *Phono: applicateurs de règles phonologiques*, CIRAL, Université Laval, Logiciel et documentation.

[8] Paradis, Cl. Deshaies, D. (1990), « Rules of stress assignement in Québec French: Evidence from perceptual data » *Language variation and change*, vol.2 pp. 135-154.

[9] Thibault, L. (1994), *Étude exploratoire du rythme en français* québécois, M.A. Dissertation, Université Laval.

[10] Santerre, L. (1991), «Incidences du trait phonologique de durée vocalique sur la prosodie du français québécois » *Proceedings of the XIIth International Congress of Phonetic Sciences*, vol. 4, pp. 254-257.

[11] Walker, D. (1984) *The pronunciation of Canadian French*, Ottawa: University of Ottawa Press.

PROSODIC ISSUES IN CZECH: AN APPLICATION IN TTS

Zdena Palková and Miroslav Ptáček
Institute of Phonetics, Prague, Czech Republic

ABSTRACT

The Czech prosody algorithm in TTS diphone synthesis is based on segmentation into stress units. It is sensitive to their positions in sentences; it creates separate sound patterns for duration, and for the F0 contour. Results confirm a hypothesis that prosodic modulation of a Czech text is well conceivable in terms of a string of linear units having characteristic sound contour, rather than as a recurrence of prominences on a neutral backdrop.

1. BACKGROUND

The purpose of research is to verify hypotheses concerning prosodic properties of Czech. Designing a plausible algorithm for the suprasegment level of the text-to-speech diphone synthesis is a means to this end, and has also a practical goal.

The description of Czech prosody can get a suitable footing in a three-tier hierarchy model of linear suprasegmental units: *syllable - stress unit* (a group of syllables having a single word stress) - *intonation unit* (a group of stress units joined by intonation). Some properties of Czech make this conceptual framework applicable at explaining general phonological issues in the prosody of Czech, and also at analysing individual text utterances, [1].

The approach adopted in our present research is also underpinned by a hypothesis that assumes as primary the very fact of segmenting the string of syllables, i.e. grouping syllables to stress units and stress units to intonation units. The search is for sound means that are instrumental in such grouping, and also for perceptual boundary signals. In other words, the linear rhythmic units do not a priori derive from prominences (word or sentential accents) as in traditional approaches. This approach responds to earlier experiments in search of the acoustic features making the nature of the word stress in Czech, [2].

(Concerning the Czech word stress, basic structural descriptions of Czech characterise it as fixed on the first syllable of words while the difference of stressed and unstressed syllables has no impact on quality or quantity of vowels.)

The key unit of the present description is the stress unit, which makes the intermediary tier, as it can reflect projections of some syllabic features, and also some features of its intonation unit. Furthermore, the unit is rather easy to mark out in Czech written texts since it has close links to the word. The necessity of getting support from phenomena definable in written texts by formal analysis follows from features of a text-to-speech production, and some trade-offs have to be done because of it, [3].

2. PROCEDURE

The following principles apply in the TTS algorithm.

a) The stress unit position is related to the text structure. Ideally, the relevant superordinate unit would be the intonation unit. Its determination in a particular text depends, however, also on syntactic and semantic information contingent on a larger context. For the purpose of automated synthesis, a "syntactic unit" is being used, which is defined as the text enclosed within two punctuation marks, in terms of Czech orthography (slightly adjusted), as the punctuation in Czech conforms to formal rules motivated largely by syntactic sentence relations.

The following additional distinctions are made within those discriminated units if necessary:

(I) - the initial position, the first stress unit in a syntactic unit

(M) - the medial position, not bordering on a following syntactic unit boundary; another stress group follows that belongs to the same superordinate unit

(F) - the position preceding the boundary signal of the syntactic unit that is not the last one in the utterance (i.e. the position preceding a comma, or a colon)

(FF) - the position preceding the boundary signal of the syntactic unit that is the last one in a finished utterance (i.e. preceding such marks as . ; ? !).

The FF-F-I-M order of preference is used when evaluating positions in the bordering versions of short texts.

b) Acoustic features are set separately for stress units of various lengths.

c) Acoustic features of stress units consist of a number of acoustic qualities, and their production algorithms have also separate designs.

No special features have been set out for initial syllables as "stress" bearers. The pattern of acoustic features spreads across the whole of stress unit. Duration modification of segments (phones or diphones) in stress units, and modification of the F0 contour, are fundamental. Modification of the dynamics is reserved for phenomena from higher levels of the text structure (emphasis, etc.); so far it has not been worked out in our automated synthesis.

Provisions have been made to allow for single acoustic features to modify the chain of stress units in various ways related to their positions within the sentence unit. E.g. a one-syllable word in the I position combines with the following word into a single unit through modifying duration but not through F0.

3. SEGMENTATION INTO STRESS UNITS

A text containing sequences of polysyllabic words submits in Czech successfully to the hypothesis that each word makes a stress unit. Variability is introduced by a one-syllable word which can stand apart, or tie to the neighbouring words. In natural speech, the solution depends on semantic and pragmatic aspects.

An algorithm has been designed to enable for such monosyllables to get connected in larger stress units based solely on the unit length and position within the syntactic unit. Tendencies found in spontaneous natural speech are made use of: The length of the produced units is 6 syllables or less, and when segmenting longer chains of syllables, the parting is either symmetrical or leaves the first portion longer. Special rules apply for single monosyllables in the I and F positions. Some exceptions have to be stored. Infrequent cases of a one-syllable word in a position "unstressed" by rule when expected to bear emphasis from its context still present a challenge.

4. DURATION PATTERNS

Duration of phones within stress units is probably influenced by the number of syllables and their types (presence or absence of a coda). Syllable breaks in Czech, however, are difficult to detect due to an abundance of consonant clusters. That is why differences in segment duration are derived in synthesis simply out of the number of phones in the stress unit.

Listening tests showed that, in perception of stress units in continuous speech, acceptable alterations in average duration of phones reflect the relationship

$$T(n) / T(m) = (m/n)^{0.12}$$

where $T(n)$ is the average phone duration in a stress unit containing n phones, and $T(m)$ is the average phone duration in a reference stress unit containing m phones. The reference number of phones has been set to 5, i.e. $m = 5$, as the stock of diphones was extracted mainly out of 5-phone words. If a relative value of 100% is set to the average phone duration in a 5-phone unit, the average phone durations get values as in tab. 1. *(See the table next page.)*

In accordance with results of listening tests, the phone durations in stress units longer than 12 phones have been exposed to no further reductions. Experiments have also shown that parameters of duration differences found out in isolated words cannot be applied. Such words reach their exponent value of 0.41 while keeping an analogous relationship between duration and a number of phones in the stress unit, [4]. The fact that larger differences in duration of phones are not acceptable in the

Tab. 1. Duration of phones related to their number in the stress unit.

Number of phones in a stress unit:											
1	2	3	4	5	6	7	8	9	10	11	12
Relative duration of a phone:											
120	111	106	102	100	98	96	94	93	92	91	90

perception of continuous speech is in agreement with the usual ranking of Czech with the syllable-timed languages.

5. PITCH PATTERNS

The F0 contour in stress units is made up separately for I+M+F and FF positions.

5.1 Stress Units in the I+M+F Positions

a) Selection of fundamental pitch patterns (i.e. F0 contours) has been based on earlier experiments which required for listeners to break up continuous flow of sound into stress units lacking any possible semantic clues [5]. Particular configurations of pitch patterns within types have been established through selection from a larger number of variants with regard to acceptability for listeners.

b) Separate pitch patterns have been set up for each stress unit of a particular number of syllables. Each type distinguishes two sets of variants (containing 2 - 6 patterns each):

(A) an off-falling contour (the last change is the fall, the pitch pattern stops at the same or lower tone related to its first syllable),

(B) an off-rising contour (the last change is the rise, the pitch pattern stops at the same or higher tone related to its first syllable).

A level contour (i.e. lacking any change of pitch) is used only with two-syllable stress units, and operates as a member of the A set.

A maximum change of pitch within a pitch pattern is 5%.

c) Rules have been accepted for chaining pitch patterns along stress unit sequences in specific texts. They consist mainly of ongoing alternations of variants from the A and B sets while the selection of a particular pitch pattern is unconstrained.

d) The phonologically relevant F position is implemented with a distinctively linked pitch pattern, usually through lowering the initial syllable of the stress unit by -2% against the closing syllable of the preceding unit.

e) Auxiliary rules have been established to solve some specific situations, e.g. for monosyllables in the F and I positions.

f) Possible excessive drop or soar of F0 in a longer chain of stress units has so far been dealt with by means of follow-up corrective rules applicable at the end of each stress unit in the FF position, i.e. after modulation in the concluded utterance have been created.

5.2 Stress Units in the Closing FF Position

a) Again, separate pitch patterns have been set up for each stress unit of a particular number of syllables. Changes in pitch have been verified by listening, and do not exceed 15%. The A set of pitch patterns (2 - 4 patterns of each length) are in force preceding the punctuation marks . ; ! and some of the ?. The B set of patterns (2 patterns of each length) apply before ?. A pitch pattern selection in questions is necessary because the F0 contour is relevant in yes-no questions in Czech. A decision has to be based on a set of rules detecting key words stored in advance.

b) The link of a stress unit in an FF position is controled by a rule sensitive to the preceding stress unit F0 contour.

6. CONCLUSIONS

The algorithm that has been worked out provides our synthesis of Czech with a prosody in a well acceptable shape. Infrequent mis-modulations arise from the inability to apply wider context-based semantic information via formal rules. Occasional corrective signals to non-standard segmentation or pitch pattern placement have to be introduced so far on an ad hoc basis.

The results seem to support a hypothesis that prosodic modulation of a stream of syllables in Czech is well conceivable in terms of a string of linear units having characteristic sound contour, rather than as a recurrence of a prominent syllable on a neutral backdrop.

Further research is now going to focus on the issue of analogous relationships within the higher intonation unit, and on reaching their formal description. It appears possible to investigate implications of such an approach in various prosodic issues in Czech and also in applications aiming at automated speech recognition.

REFERENCES

[1] Palková, Z. (1994), *Fonetika a fonologie češtiny,* Praha, Karolinum 1994.

[2] Janota, P. and Palková, Z. (1974), "The auditory evaluation of stress under the influence of context", *Phonetica Pragensia IV: Acta Universitas Carolinae,* Praha, pp. 29-59.

[3] Palková, Z. and Ptáček, M. (1994), "Ein Beitrag zur Intonation in der Diphonsynthese", *Phonetica Pragensia VIII: Acta Universitas Carolinae,* Praha, pp. 61-74.

[4] Palková, Z. and Ptáček, M. (1994), "Der Sprechtakt als eine rhythmische Einheit in der Diphonsynthese der tschechischen Sprache", *Speech Processing: 4th Czech-German Workshop,* Prague, p. 23.

[5] Palková, Z. (1987), "Intonatorische Merkmale in der Perzeption der Wortgrenzen im Satz", *Proceedings of the XIth International Congress of Phonetic Sciences,* vol. 1, Tallinn, pp. 296-299.

ACOUSTIC PROPERTIES OF DISFLUENT REPETITIONS

E. E. Shriberg
Institute for Perception Research (IPO), Eindhoven, The Netherlands
and SRI International, Menlo Park, CA, USA

ABSTRACT

Acoustic properties of disfluent repetitions are examined to investigate two proposed functions of repeating. Repeating may serve as a filler while hesitating; alternatively, repeating may function to bridge a gap when speech resumes after a break. Classification of tokens based on pause patterns reveals that: (1) most cases fit the bridging function; and (2) duration and F0 properties support the hypothesis of two distinct types, as well as the proposed associated functions.

INTRODUCTION

Speakers often repeat words in spontaneous speech, resulting in lexical disfluencies such as that shown in Figure 1.

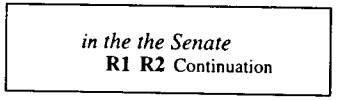


Figure 1. Example of a disfluent repetition and terminology.

Little is known, however, about *why* speakers utter the repeated instance (R2).

Heike [1] suggested two alternative functions of repeating, which could be distinguished based on the presence of an unfilled pause following R2. Possible surface patterns for the hypothesized functions are summarized in Figure 2.

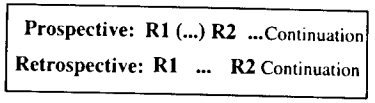


Figure 2. Surface patterns for proposed functions. "..."=pause; "( )"=optional.

He termed cases in which R2 was followed by a pause ***prospective repeats***, suggesting such repeats serve a stalling function, to hold the floor during hesitation. Cases in which R2 was not followed by a pause (but preceded by a pause) were termed ***retrospective repeats,*** and were proposed to function as bridging devices to connect a continuation with preceding material after a break in fluency.

Although these proposed functions are reasonable theoretical possibilities, in practice there is little if any empirical evidence to support the claim that there are two types of repeats. For purposes of this paper, we will assume that there *are* two different functions of repeating, and that the functions can be distinguished on the basis of whether or not R2 is followed by an unfilled pause before the continuation. Leaving the issue of function aside, the terms "prospective" and "retrospective" will be used simply to refer to the classification of the repetition based on its surface pause characteristics.

Given these assumptions as a starting point, this paper seeks to answer two questions about the different types of repeats. First, are both types actually found in spontaneous speech data, and at what relative rates? Second, if we look more closely at acoustic properties of repetitions, can we find characteristics other than unfilled pauses that pattern differently for the two types?

METHOD

Single-word disfluent repetitions were extracted from the speech of six speakers (three male, three female) in the SWITCHBOARD corpus of telephone conversations [2]. In this corpus, speakers conversed with an unfamiliar partner on a chosen topic. Despite the somewhat contrived task, conversations were rated as highly natural-sounding by transcribers. Selection of repetitions was limited to cases with no other disfluency either between or directly following the repetition.

Hand-labeling of 242 such cases was conducted using the GIPOS speech analysis package developed at IPO. For each example, five time values were recorded: the onset and offset of R1, the onset and offset of R2, and the onset of the continuation. For examples with adequate F0 tracks for both R1 and R2, and in which both R1 and R2 showed a roughly linear F0 trajectory, the first good F0 and last good F0 of R1 and of R2 were also recorded.

RESULTS AND DISCUSSION

Frequency of types

Figure 3 shows the frequency of types (as classified based on the presence of a pause following R2). As shown, both

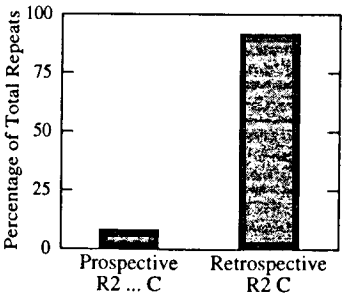


Figure 3. Relative frequency of types. "R2"=repeated instance, "C"=continuation, "..."= unfilled pause.

patterns occur in the speech data examined. However, the clear majority of cases were classified as retrospective, i.e. cases in which R2 is hypothesized to serve a bridging rather than stalling function.

Acoustic properties of types

To address the second question posed in the introduction, duration and F0 properties of R1 and R2 were examined in an aim to provide evidence other than simple pauses to support Heike's claims.

Duration. In Figure 4, the duration of R1 is plotted against the duration of R2 for all tokens (over all words and speakers). Different symbols denote the two different repeat types. The equivalence line (y=x) is indicated for reference.

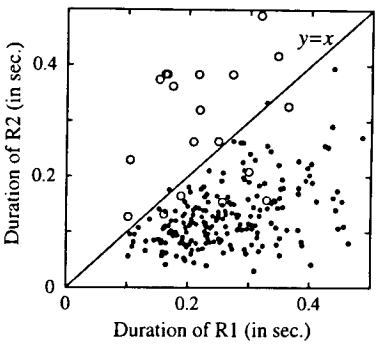


Figure 4. Duration of R1 versus R2 for all tokens; ○ = prospective repeat, ● = retrospective repeat.

As shown, there is a clear difference between the two types. Data for prospective repeats occur both above and below the equivalence line. The data for retrospective repeats, however, are nearly all below the equivalence line, indicating that R1 is systematically longer than R2.

In order to interpret these data, however, we must know whether R1 is lengthened, or R2 shortened (O'Shaughnessy suggested that both effects occur [3]). To address this issue, a small study controlled for speaker and word was conducted. Durations for R1, R2, and unrepeated (fluent) instances of the word "the" were compared for the two speakers with the largest amount of data. The fluent examples were chosen from those conversations which also contained the repeated instances. For speaker 1, 19 repeated tokens and 40 unrepeated tokens were obtained; for speaker 2, 12 repeated and 33 unrepeated tokens were obtained. Results are shown in Figure 5.

Despite the small sample sizes, there is a significant difference between R1 and the other two conditions, as can be inferred from the error bars. Also, importantly, R2 does not appear to be shortened since it appears actually slightly longer than unrepeated instances.

Thus, for retrospective repeats, there is lengthening at R1 and no lengthening at R2; this is consistent with the bridging

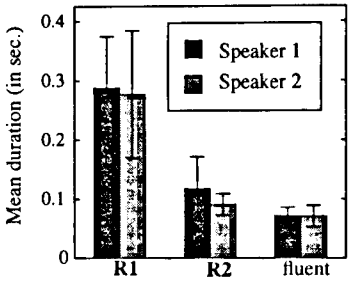


Figure 5. Mean duration of R1, R2, and unrepeated instances of "the" for two speakers.

function proposed by Heike for such cases, since R2 resembles a fluent resumption.

Given these results, inspection of absolute durations in Figure 4 also suggests that R2 is lengthened for most prospective repeats, consistent with the proposal that in these cases R2 functions as a hesitation device.

Fundamental frequency. A difference between prospective and retrospective repeats was also found in F0 properties of R1 and R2. When the four measured F0 points described in the method section were plotted at equal intervals (i.e. not taking into account the duration of the words), results showed roughly parallel trends at different F0 ranges, although this requires modifying the linear scale (an appropriate scaling model is described in [4]). Thus, a representative picture of the relationships between these four points can be conveyed by plotting the mean values for each speaker. Such values are plotted for retrospective repeats in Figure 6. Values are shown separately for males and females in order to display results on appropriate scales.

As shown, both words (in nearly all cases words were unaccented) fall in F0. A consistent tendency across speakers is that the onset of R2 is reset to a value about equal to that of the onset of R1. It is also notable that R2 falls in F0 like R1,

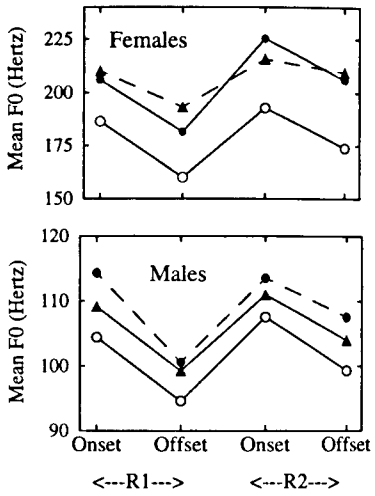


Figure 6. Mean onset and offset F0 of R1 and R2 in retrospective repeats, by speaker.

but not to as low a value; this is probably due to the much shorter duration of R2.

It was not possible to obtain data for all speakers for the set of prospective repeats due to low sample size. However, in Figure 7, results are shown for one female speaker. The data for the retrospective repeats from the same speaker (from Figure 6) have been replotted in Figure 7 for comparison.

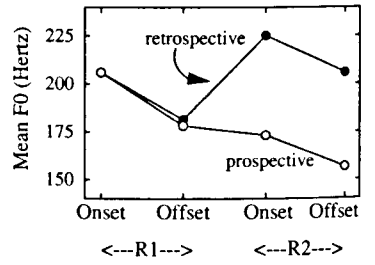


Figure 7. Mean onset and offset F0 of R1 and R2 in retrospective and prospective repeats for a single speaker.

As shown, the two types show different patterns. Unlike retrospective repeats, prospective repeats show an R2 that continues to fall in F0 from R1. Similar results were obtained for other speakers. The continuous fall in F0 for the prospective repeats is consistent with an observation for filled pauses, which are also proposed to serve a stalling function: it was noted informally in past work that sequential filled pauses (e.g. "uh uh uh") showed successively lower starting F0 values.

Although some prospective repeats showed a reset, retrospective repeats rarely lacked the reset. The reset for the retrospective repeats is consistent with findings by Levelt and Cutler [5], who observed that repairs tend to be uttered at the same F0 values as the material they replace.

SUMMARY AND FUTURE WORK

This research found that when repetitions were classified based on pause characteristics, the majority of cases showed no pause between R2 and the continuation. Such cases are consistent with a bridging, rather than a stalling function.

Furthermore, analyses revealed that the classification of tokens based on pauses correlated with durational and F0 properties of R1 and R2. This result adds weight to the the proposal that there are two different types of repeats. In addition, the duration and F0 properties pattern in ways that are consistent with Heike's proposed functions.

An important next step in this line of research is to investigate the question of function directly, for example by conducting controlled elicitation or perceptual experiments. Goals for such future work include gaining knowledge about factors (e.g., syntactic, prosodic, task-related, speaker-related) that influence the production of repeats, as well as gaining an understanding of how repetitions function both for speakers and listeners.

ACKNOWLEDGEMENTS

The author gratefully acknowledges Marc Swerts for valuable comments and discussion. This work was supported by an NSF-NATO postdoctoral fellowship. It is also part of ongoing research funded by the Advanced Research Projects Agency under NSF Grant IRI-9314961, and by NSF Grant IRI-8905249. The views and conclusions contained in this document are those of the author and should not be interpreted as reflecting the official policies of the funding agencies.

REFERENCES

[1] Heike, A.E. (1981). A content-processing view of hesitation phenomena. *Language and Speech, 24* (2), pp. 147-160.

[2] Godfrey, J.J., Holliman, E.C. & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proc. ICASSP*, pp. 517-520.

[3] O'Shaughnessy (1993). Analysis and automatic recognition of false starts in spontaneous speech. *Proc. ICASSP*, pp. 521-524.

[4] Shriberg, E.E. (1994). Preliminaries to a theory of speech disfluencies. Unpublished Ph.D. thesis, University of California at Berkeley.

[5] Levelt, W.J.M. & Cutler, A. (1983). Prosodic marking in speech repair. *Journal of Semantics, 2* (2), pp. 205-217.

THE RELATIONSHIP BETWEEN RHYTHM AND THE TONE SANDHI DOMAIN IN MANDARIN

J. Wang and R.H. Mannell
Speech, Hearing and Language Research Centre
Macquarie University, Sydney, Australia

ABSTRACT

In this paper, we develop the notion that the relationship between prosody and the tone sandhi domain in Mandarin is rhythm-based, determined by the strong rhythmic tendency to group words into disyllabic and trisyllabic units. Three experiments were conducted to test the relationship between rhythmic grouping and tone sandhi, as well as the effect of speech rate and focus position.

INTRODUCTION

Tone sandhi (TS) of the 3rd tone in Mandarin has long been of phonological and phonetic interest particularly because of its relationship to other prosodic features, and to syntactic structures. With the aim of developing rules to predict the TS domain, phonological studies have focused on syntactic, metrical and prosodic approaches and there is a general consensus that TS domain is determined by prosodic structures [1] [4] [7]. Acoustic phonetic studies have also been carried out on tones and tone sandhi in different tonal and prosodic contexts [3].

Prosodic approaches to TS, have tended to suggest that the TS domain is based on the prosodic grouping of words, with the units of grouping being disyllabic feet, super feet, and phrases [4], but such approaches (eg. *Foot Formation Rules*), have limited predictive power as little is known about how syllables and words are grouped in different prosodic contexts.

In this study we examine the following hypotheses: (1) the prosodic conditioning of TS is rhythm-based and is strongly related to the tendency of speakers to group words into disyllabic and trisyllabic units; (2) in most situations, word-level rhythmic grouping is relatively stable, although changing speech rate and focus position could, in certain circumstances, affect rhythmic structures, leading to variations in the TS domain.

RHYTHMIC GROUPING AND THE TONE SANDHI DOMAIN

In connected Mandarin speech, syllables and words are grouped into rhythmic units. The predominant rhythmic tendency is to use disyllabic constituents, either lexical words or the combination of two monosyllabic words. This tendency implies that monosyllabic words tend not to stand as independent prosodic (rhythmic) units. As a result, monosyllabic lexical words are normally grouped with neighbouring words to form basic rhythmic units. In fluent, natural speech, priority is given to grouping monosyllabic words with neighbouring monosyllabic words, then disyllabic words and, finally, polysyllabic words with more than two syllables — unless there is good reason not to, such as pausing. The constituents resulting from this type of grouping are basically disyllabic or trisyllabic independent prosodic units. We prefer to call them prosodic words (PW), rather than feet, since the term "foot" is traditionally related to the alternation of stressed and unstressed syllables. Lexical words (rather than syllables) are the basic elements of this type of rhythmic grouping, and there are three main factors determining the formation of prosodic words: a syllable count of each lexical word; a syllable count of each of its neighbours; and the speaker's desire to form a balanced rhythmic structure. A syllable count of each word determines if it is obligatory to group the word with neighbouring words, while syllable counts of neighbouring words influence the grouping priority (the way words are grouped to form balanced rhythmic structures). In fluent speech with a normal speech rate, for example, the rhythmic structures 2/2, 3/2, 3/3 are likely in the following sentences:

a. laoli / **mai** gou.
b. laoli **mai** / xiaogou.
c. laoli **mai** / xiaomugou.

Here, the monosyllable *mai* is grouped with either the ensuing word *gou* or the preceding word *laoli*, because of the speaker's desire to form balanced rhythmic structures, rather than being determined by syntactic branch directions or semantically related elements. In practice, the monosyllabic verb *mai* can be grouped with either *xiaogou* or *laoli* in example *b.* without affecting the rhythmic balance of the sentence, although grouping it with *laoli* is usually preferred. However, for sentence c, it is virtually mandatory in normal speech to group *mai* with *laoli*, as 3/3 is a more balanced structure than 2/4.

Tone sandhi for the 3rd tone in Mandarin is motivated by dissimilating adjacent low tones and tends to operate within units of speech planning [1]. Our research indicates that speech planning units are created from rhythmic groupings rather than *syntactically or semantically related elements.* Although Foot Formation Rules work well in most situations, and were originally designed to *modify syntactic structure and produce rhythmically balanced structures for TS to operate on* [4], limitations arise because they refer to conditions for syntactic branch direction, rather than to rhythmic groupings.

If the TS domain is identical to the rhythmic grouping of words, as we have previously argued, then factors which influence rhythmic groupings must also affect the TS domain (such as speech rate, focus stress and pauses), since the surface TS patterns should reflect the underlying groupings. Therefore, close examination of the effects of these factors on the TS domain should further clarify how words are rhythmically grouped in Mandarin speech. There has been a lot phonological discussion on the basic TS domain and variations in the TS domain, due to speech rate and contrastive stress [4] [7], but there has been little acoustic investigation carried out to date [6]. In this study, three related experiments were carried out to examine, both aurally and acoustically, the relationship between rhythmic groupings and the TS domain, as well as the effect of speech rates and focus position on rhythmic groupings and TS domain.

MATERIALS AND PROCEDURE

12 sentences, comprising of only 3rd tone syllables, were read by four female Mandarin speakers from Beijing (XY, LP, LL, WJ). The subjects were asked to read the sentences naturally with three different speech rates, two or three times initially, and then according to specific contexts that required changing contrastive stress positions across syllables and words. All the sentences were recorded in studio conditions using DATs and the speech samples were then digitised onto a Sun computer at a sample rate of 20 kHz, and manually segmented and labelled using the Waves program. Relevant duration data and pitch contours were extracted and calculated using the mu+ system [2].

a1 wo xiang mai gou.
a2 wo xiang mai xiaogou.
a3 wo xiang mai xiaomugou.
a4 wo mai gou.
a5 wo mai xiaogou.
a6 wo mai xiaomugou.
b1 laoli xiang mai gou.
b2 laoli xiang mai xiaogou.
b3 laoli xiang mai xiaomugou.
b4 laoli mai gou.
b5 laoli mai xiaogou.
b6 laoli mai xiaomugou.

RESULT 1: BASIC GROUPING

In this experiment, 12 sentences were read, fluently and naturally, three times by each of the subjects. The following table

shows the the TS domain. The results are relatively consistent, with identical PWs used by each of the four subjects. An analysis was carried out on the TS patterns with each syllable's tone being identified aurally. Rhythmic structures (prosodic word boundaries) were determined according to the prosodic rules outlined earlier. In the following table, 'R' refers to TS occurring with the syllable's tone being transformed into a rising tone, while 'L' refers to the syllable keeping its underlying tone. The slash '/' indicates PW boundaries.

a1 RL/RL	b1 RL/RRL
a2 RRL/RL	b2 RL/RL/RL
a3 RRL/RRL	b3 RL/RL/RRL
a4 RRL	b4 RL/RL
a5 RL/RL	b5 RRL/RL
a6 RL/RRL	b6 RRL/RRL

Each sentence was read three times by each of the subjects and some idiosyncratic variations of TS were observed. These variations can be categorised into three types: Type I — using alternate TS patterns, such as LRL instead of RRL in trisyllabic PWs with 1+2 syntactic structures, like *(xiao(mu(gou)))* in b3; Type II — joining two PWs together, so that RRRL occurs instead of RLRL, as in b4; Type III — grouping monosyllabic words in different ways in certain contexts, such as in b1 and b5 where the rhythm could be either 2/3 or 3/2. Our results indicate that there are conventional rhythmic structures for each of the sentences but there is also a certain degree of individual freedom.

RESULT 2: SPEECH RATE

We also asked the subjects to read each of the sample sentences at different speech rates for purposes of testing the effect of speech rates on the TS domain and PW grouping. The speech rates (syllables per second) are, as one would expect, slightly different across the four subjects.

The following table gives the average speech rates for each of the four speakers at slow, medium and fast speech rates.

	XY	LP	LL	WJ
medium	4.5	4.1	4.4	4.8
fast	5.4	6.5	5.2	6.2
slow	3.2	3.1	3.3	3.7

We found that TS domains in the majority of the sentences were reasonably consistent for the various speech rates. As was the case in the previous section, three types of variations were observed. Type I occurred more frequently in slow speech, while type II was more common in fast speech. Type III had no definite correlation with speech rate. While it appears that more tonal transformations occur in fast speech than in slow speech, either through using alternate TS patterns or by grouping two PWs together, the basic rhythmic groupings remain relatively stable, regardless of speech rate.

RESULT 3: FOCUS POSITION

The position of the accent has been reported to influence the TS domain as TS always starts from accented syllables [7]. In this experiment, the location of contrastive accents is designed to change across syllables and words according to different contexts, and subjects were asked to read a dialogue in which they could stress appropriate syllables or words.

Q: Ni xiang mai xiaogou ma?
(Do you want to sell that small dog?)
A: Bu, wo xiang **mai** xiaogou.
(No, I want to buy a small dog.)

There are 54 contexts for the 12 test sentences in which the location of contrastive accent varies. Our results showed that changes in TS domain do occur according to the position of accented words or syllables in certain contexts, but that it is extremely difficult to determine whether these changes are accent-conditioned. First, TS domains in most of the sentences are quite stable, regardless of the location of accented syllables. For example, TS domain in b4, b5, b6 is always 2/2, 3/2, 3/3, regardless of whether *mai* is accented or not.

We then looked at prosodic contexts which caused rhythmic regrouping and found there were only two examples that consistently occurred across each of the four subjects. The normal TS domain of b2 is RL/RRL, but RRL/RL when *mai* is accented; while in a1 it is normally RL/RL, but L/RRL when *xiang* is accented. If TS domain changes are caused by different planning groupings, it seems to be associated with not only the position of accented syllables but also PW structures. It was found that rhythmic regroupings only occurred in sentences with more than two successive monosyllabic words and accented-related regrouping never split polysyllabic lexical words into two separate PWs. For example, *xiang* can be grouped with *laoli* in b4, if it is necessary to accent *mai* using a full rising tone, but this is not the case in b5 when *xiao* is accented. There is a basic distinction between PWs consisting of only one lexical word and those with one or more words. The latter have a certain flexibility to be grouped in different ways, which is not only illustrated in the TS domain, but also in the acoustic data. For example, the syllable duration of *gou* (see the table below) not only varies according to whether it is accented or unaccented but also according to whether it is a monosyllabic word or the last syllable of a polysyllabic word. (In this table, 1 is *gou* in a monosyllabic, 2 in a disyllabic, and 3 in a trisyllabic word. "a1" etc. are sentence numbers. "238" etc. are durations in ms).

	unaccented	accented
1 (a1, a4, b1, b4)	238	354
2 (a2, a5, b2, b5)	220	297
3 (a3, a6, b3, b6)	220	295

When *gou* is accented as a monosyllabic word, the degree of lengthening is significantly increased. This implies that monosyllabic words, as meaningful units, have a relatively loose relationship with neighbouring words within PWs. This may explain why accented words can affect rhythmic structures and TS domains.

CONCLUSION

This study has attempted to combine phonological and phonetic approaches to the TS domain within a rhythmic framework. We have found that basic rhythmic structures are relatively stable, but that certain spontaneous factors can influence grouping. Relatively stable rhythmic structures would appear, therefore, to be the logical starting point for developing phonological rules for predicting the TS domain, but the types of spontaneous factors which can occur make the perfection of these rules extremely difficult. This rhythmic framework has been tentatively applied in the hierarchical labelling of rhythmic units in a Mandarin speech database aimed at establishing a prosodic model for Mandarin speech synthesis [5]. However, at this stage, further acoustic investigation of rhythmic grouping in Mandarin speech is still required.

REFERENCES

[1] Chen, M. (1990), "What must phonology know about syntax", in *The Phonology-Syntax Connection*, S. Inkelao and D.Zek (eds.), U.Chicago Press, 19-46.
[2] Harrington, J., Cassidy, S., Fletcher, J., McVeigh, A. (1993), "The mu+ system for corpus based speech research", *Computer Speech and Language*, 7, 305-331.
[3] Speer, S.R., Shih, C.L., Slowiaczek, M.L. (1989), "Prosodic structure in language understanding: evidence from tone sandhi in Mandarin", *Language and Speech*, 32(4), 337-354.
[4] Shih, C. L. (1986), *The prosodic domain of tone sandhi in Chinese*, PhD dissertation, Univ.of California, San Diego.
[5] Wang, J. (1994), "Syllable duration in Mandarin", *Proc. 5th Australian Int. Conf. Speech Science and Technology*, 322-327.
[6] Zee, E. (1980), "A spectrographic investigation of Mandarin tone sandhi", *UCLA Working Papers Phonetics*, 98-116.
[7] Zhang, Z. S. (1988), *Tone and tone sandhi in Chinese*, PhD dis., Ohio State U.

ACADEMIC ENGLISH SPEECH INFORMATION CONTINUUM AND TEMPORAL STRUCTURE

N.S.Yelkina, Kiev Pedagogical Linguistic University, Kiev, Ukraine

ABSTRACT

This study is an attempt to reveal the function of the temporal component of intonation in the actualization of the semantic ctructure of an oral academic discourse. Under examination are such temporal parameters as speech rate and pause distribution.

BASIC ASSUMPTIONS

It is argued that there is a certain correlation between the temporal and semantic structure of oral discourse and in particular it is asserted that tempo and pause distribution are major prosodic markers of the functional sentence perspective: a delimitative pause between a rheme and a theme and the slowing down of tempo in the rheme-containing syntagms /sense-groups/ are considered to be its main attributes. But in speaking as opposed to reading we face a whole range of psycholinguistic phonetic phenomena related to speech coding process which tend to disrupt this pattern. Moreover, the semantic structure of an individual utterance is conditioned by the semantic structure of the whole phonopassage in which it occurs, and ultimately, by that of the whole discourse which may be another reason for variations in tempo and pause distribution in an utterance.

METHOD AND EXPERIMENTAL CORPUS /EC/

The EC comprises 3 academic lectures delivered in a sound-proof studio by English speakers who are teachers by profession and have had extensive experience of public academic speaking. All the lectures had been delivered previously to a students' audience ; during the recording the speakers were told to make very little use of notes if any at all. the experiment included 3 types of analysis: semantic, auditory and acoustic.

SEMANTIC ANALYSIS. DATA OBTAINED

We followed the semantic analysis developed by T.M.Dridze /1/, which is based on the concept that any discourse can be viewed as a hierarchy of semantic units /predications/ of varying semantic value, among which the 1st & 2nd order predications have the highest semantic status since they include such informative elements as the main aim of discourse, one or several main propositions, their explication and situation evaluation. The 3d order predication os composed of illustrations to the 1st & 2nd order predication elements, while the 4th order predication represents a semantic background to the main aim of discourse.

In Fig.1 is shown the semantic macrostructure of Lecture 1 ("British Accents"). As is seen the aim of the lecture falls into a number of autonomous sub-aims, each one referring to a specific accent. For our purposes we focused only on the thematic fragment dealing with RP. As is evident Sub-aim 1 is expressed by 7 propositions (A-1a ... A-1g). Some of the propositions may be semantically amplified by explanatory (A-2a,A-2d,A-2f,A-2g), evaluative (A-3d), illustrative (B-1.1b,B-1.2f) elements, whereas the others do not receive any semantic amplification. A proposition with its semantic amplifiers, if any, forms a complex semantic unit which in discourse syntagmatics is realized as a phonosemantic complex /PSC/ possessing both semantic and prosodic integrity and varying in length from 1 to 4 phonopassages, the main proposition forming its nucleus (see PSC 1 & PSC 2

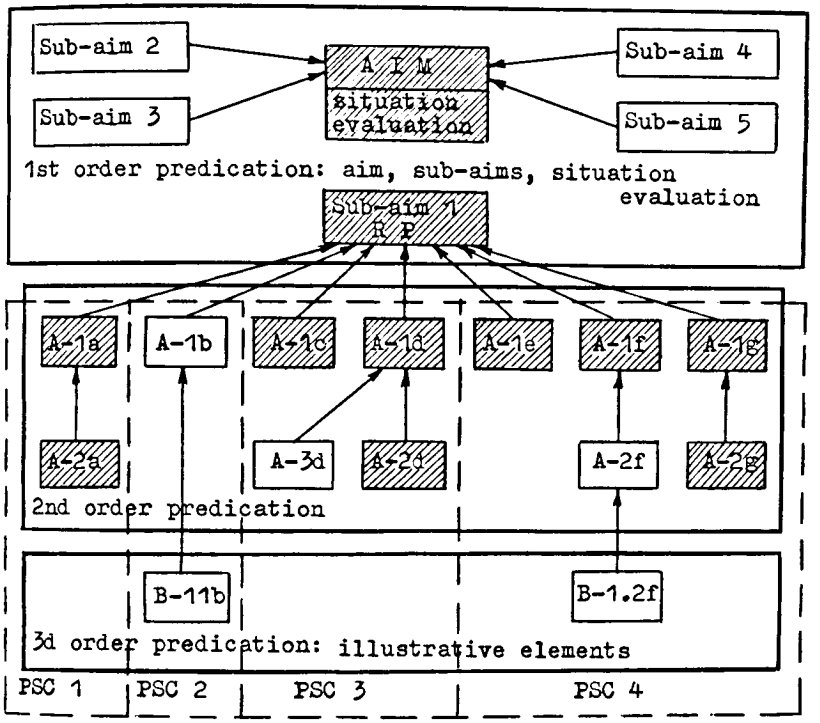


Figure 1. Semantic macrostructure of Lecture 1.

in Fig.1). A PSC may also be centered around a composite nucleus of several propositions if they are joined together in 1 phonopassage and thus make an indivisible phonetic unit (see PSC 3 & PSC 4). So the structural hierarchy of spoken discourse is: phonosemantic complexes - phonopassages - phrases - sense-groups (syntagms).

As is obvious this method helps to reveal the paradigmatic relations among predication elements of varying orders and may serve as a basis for discourse structural typology.

The EC is represented by 6 phonopassages composed exclusively of the 1st & 2nd order predications. (In Fig.1 the elements in question from Lecture 1 are shaded.) In each phonopassage theme-containing syntagms /TSC/ and rheme-containing syntagms /RSC/ are determined and subjected to auditory and acoustic analysis.

ACOUSTIC ANALYSIS. DATA OBTAINED

Under examination are:

1.general rate of speech (articulation rate & pauses)
2.location of temporal extremums
3.pause distribution.

The number of syntagms analyzed is 92 (the ratio TSC/RSC=1). As shown in Table 1 42% of RSC are pronounced at faster speech rate than the preceding TSC and in another 8% there is very little or

no change at all. At the same time in 43% of TSC a decrease of speech rate is observed as compared to the preceding RSC and in 9% of instances the rate remains stable.

Table 1.

Type of syntagm	General rate of speech		
	↓	↑	→
TCS	43%	48%	9%
RCS	50%	42%	8%

Moreover, 54% of all the temporal minimums in the EC fall on the TSC, while 46% of all the temporal maximums in the EC appear in the RCS.

This fact contradicts the well-known assertion that speech rate variation in an utterance is determined by its semantic structure, namely, speech rate usually slows down on the rheme and accelerates on the theme. We maintain that the reason for these speech rate fluctuations lies, on the one hand, in the distribution of pauses in speech flow and, on the other, in the integrative function of speech rate.

On the perceptive level 3 categories of pauses are identified: syntactical, emphatic and hesitation pauses. Oral academic discourse conveys both intellectual and volitional information, therefore the ample use of all sorts of phonation breaks for the sake of emphasis is relevant here. In the experimental data the following types of communicative phonation breaks used for this purpose are determined:

1.Micropauses and glottal stops (25 - 110 msec); occur exclusively in RCS, or occasionally in the initial syntagms of a phonopassage, are not normally perceived and serve as word boundary markers in RCS. Their function is to provide the listener with additional phonological signals which can help him in decoding the message.

e.g. ... the ˈcontent ˈshould | inˈclude||| a ˈ central ʔˈelement} ˈof{ ˈethics||.

2.Series of final delimitative pauses (75 - 290 msec); occur in one of the final syntagms of a PSC breaking it into a number of rhythmical groups thus creating staccato rhythm. Their function is to signal to the listener that the speaker has finished a certain theme (PSC) and is going to pass over to the next one.

e.g. ... my ˈaccent could disˈguise that very ˈwell | and ˈpeople would ˈnot | deˈtect that | ˈfact |||.

Interestingly, the similar phenomenon was also detected in Russian discourse, its function however was not explained /2/.

3.Rhetorical pauses (50 - 745 msec); of their total number registered in the EC 74% appear in RCS. Very often they are placed between a form-word and the following lexical word so as to draw the listener's attention to the postpausal fragment. In 50% of instances the preceding form-word undergoes an emphatic lengthening.

e.g. I suˈppose that the | ˈway that you could sum ˈup | reˈceived proˈnunciˈation| ...

The number of communicative phonation breaks is twice as high as the number of hesitation pauses (30% versus 15%) but the latter tend to be longer. They last 64 - 2226 msec and in 58% of cases are preceded by segmental lengthening.

e.g. ... ˈand | as I ˈsaid | ...

The experimental data in dicate that communicative phonation breaks and hesitation pauses have a different distribution pattern: the former tend to appeare in RCS (80% of their total number), whereas the latter concentrate mostly in less informative TCS (81% of their number). The overall pattern of pause distribution for each lecture and the average data are shown in Fig.2.

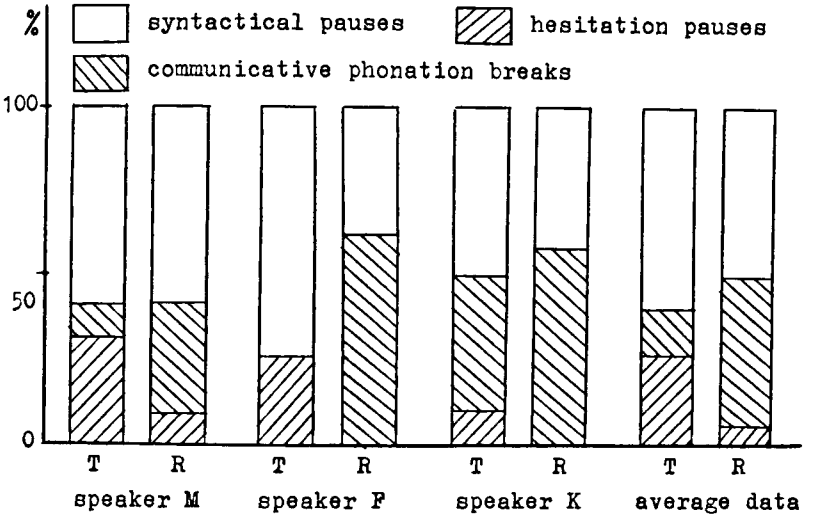


Figure 2. Pause distribution for each lecture and average data.

Thus the slowing down of speech rate in TCS may be attributed to the high incidence of hesitation pauses in them, the more so as hesitation pauses are usually accompanied by the overall decrease of tempo in a syntagm. Discourse temporal model which allows for hesitation pauses to occure predominantly in less informative TCS and is characterised by the distribution of communicative phonation breaks in more informative RCS is evidently optimal from the listener's point of view as it stimulates rather than hinders the decoding process.

IT has also been discovered that 64% of utterance-final syntagms are pronounced at a faster rate of speech that their immediate precontext. It is assumed that they mark a close retrogressive semantic link of the following utterance with the preceding one. In our case when rhemes are mostly in preposition (TCS - RCS), the temporal maximums fall on RCS.

CONCLUSION

Thus, the temporal structure of an individual utterance within a phonopassage/PSC is conditioned by 2 factors: semantic and psycholinguistic. The former manifests itself in the fact that speech rate within an utterance is determined not only by the semantic structure of an utterance but also by that of the whole phonopassage/PSC. The latter is associated with a high prevalence of hesitation pauses in less informative utterance segments. Apparently, the model described may be held as the temporal invariant of prepared academic speaking.

REFERANCES

1.T.M.Dridze (1980), Language and Social Psychology, M., p.224

2.T.M.Nikolayeva (1977), Phrasal Intonation of Slavic Languages, M., p278.

THE VOWEL SYSTEM OF ITALIAN CONNECTED SPEECH

Federico Albano Leoni, Francesco Cutugno, Renata Savy
CIRASS - Università di Napoli "Federico II"

ABSTRACT

A research on the spectro-acoustic features of Italian vowels in connected speech will be presented. The material used for our analyses was recorded from four regional TV news bulletins. 6400 vowels were analysed (40 speakers uttering 10 tokens of 16 vowel types).

We will present some results relative to the male half of the corpus. We will focus on some aspects such as reduction and centralization, overlaps between adjacent vowels, vowels duration.

All our data are available via Internet, see note in the last page after References for details.

INTRODUCTION

The earliest spectroacoustic description of the Italian vowel system was made by Ferrero in 1968 [1].

The data presented at that time were obtained from a corpus of vowels pronounced in isolation by 20 speakers native of northern regions of Italy. All the following attempts to describe the Italian vowel system suffered from the same limitations [2], [3], [4].

For the present work we decided, on the contrary, to create a phonetic database based on recordings of TV news bulletins (see also [5], [6]).

In our view this kind of speech material represents an example of standard Italian of middle-high level and, at the same time, an example of connected speech which, although partly read by the journalists, can be seen as "spontaneous".

METHODS

The total corpus consists of 6400 items taken from recordings of regional bulletins broadcast by the national TV company (RAI) in four regions of Italy: Lombardy, Tuscany, Latium and Campania. For each region 10 speakers were chosen (5 males and 5 females); from each of the total 40 speakers 10 tokens were used of 16 different vowel types (stressed [i, e, ɛ, a, ɔ, o, u], unstressed non final [i, e, a, o, u], unstressed final [i, e, a, o]). The tokens were extracted from "full words" (nouns, adjectives, verbs, adverbs) excluding diphthongs. We organized our data into a database containing the following information:

-speaker id.;
-word uttered;
-stress conditions and position of the vowel (stressed, unstressed, final);
-preceding and following segments;
-preceding and following vowels;
-syllable structure (open or closed);
-total vowel duration;
-maximum pitch within the vowel;
-maximum energy within the vowel;
-f1, f2, f3 values measured from average FFT spectrum of the central portion (second third) of the vowel.

Presently (April 95), only data from the male speakers portion of the corpus are completely available, while the analysis of the female portion is going to be completed within a few months.

RESULTS

General results

The average and σ values for f1 and f2 are presented in Table 1.

	i		e		ɛ		a		ɔ		o		u		
	f1	f2	f1	f2	f1	f2	f1	f2	f1	f2	f1	f2	f1	f2	
stressed	273	2234	375	2037	493	1855	702	1488	536	1043	404	983	307	895	mean
	54	173	62	205	76	163	91	112	66	147	62	164	57	145	σ
unstress.	283	2099	410	1775			564	1454			433	1091	315	1013	mean
	54	198	77	217			94	176			79	167	65	213	σ
final	294	2106	435	1790			567	1497			438	1134			mean
	54	212	94	176			85	142			77	172			σ

*Table 1. Mean values and σ of f1 and f2 (Hertz) for **stressed, unstressed** non final, and unstressed **final** vowels.*

In Figures 1, 2 and 3 the data of Table 1 are presented in terms of f1/f2 diagrams on a Hertz linear scale. In these figures each ellipse defines the distribution area of a vowel. The coordinates of the ellipse centers are the mean values of f1 and f2 while the axes are±1σ.

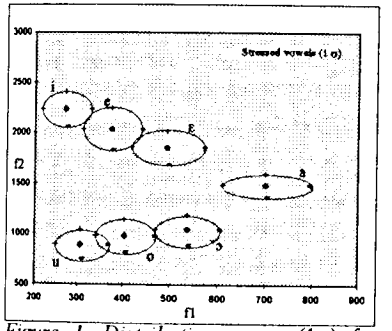


Figure 1. Distribution areas (1σ) for Italian stressed vowels.

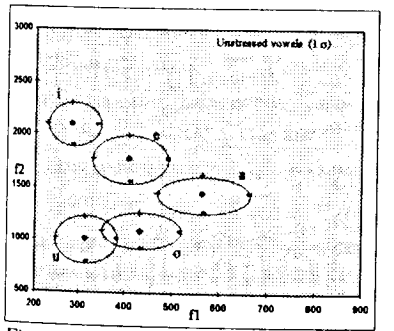


Figure 2. Distribution areas (1σ) for Italian unstressed non final vowels.

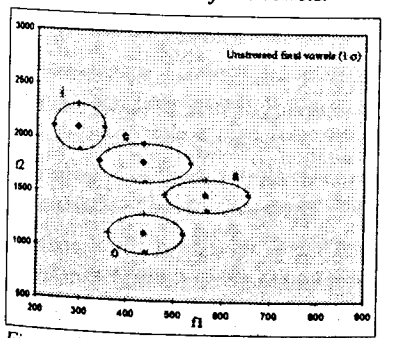


Figure 3. Distribution areas (1σ) for Italian unstressed final vowels.

Overlap of areas and centralization

As is widely known, the ellipses shown in Figures 1, 2 and 3 define a portion of the f1/f2 plane which includes about 67% of the vowels belonging to the given category: in other words this kind of representation gives an immediate idea of the most probable values for f1 and f2 for each vowel type. Observing he data distribution more closely, the following phenomena become clear:

1) within each diagram, data coming from adjacent vowel categories are strongly overlapped; as an example of this statement Fig. 4 shows the same data as Fig. 2 but the ellipse axes are now ±2σ, corresponding to the delimitation of a portion of the f1/f2 plane containing about 95% of the data of each vowel type. In this graph the overlaps between the ellipses are dramatic. Similar results would be obtained applying the same procedure to the data of Figures 1 and 3.

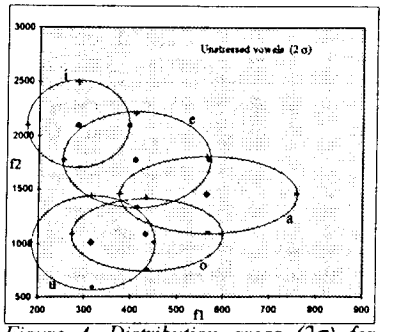


Figure 4. Distribution areas (2σ) for Italian unstressed non final vowels.

2) the comparison of mean values for similar vowel categories in different stress conditions indicates a tendency to the centralization of all vowels when unstressed. In Figure 5 the mean values of stressed vowels are compared those obtained in the unstressed condition. As in Italian the stressed vowel couples [e, ɛ] and [ɔ, o] are reduced respectively to [e] and [o] when unstressed, we compared the unstressed data with the average between the two corresponding stressed vowels. Each arrow-ended line in the graph corresponds to the 'direction' of this centralization. These lines tend to meet in a very small area. The little square in the middle of Figure 5 indicates the center point of this area.

The ideal ellipse defining such area has an f1 = 332±24 Hz and an f2 = 1350±46 Hz.

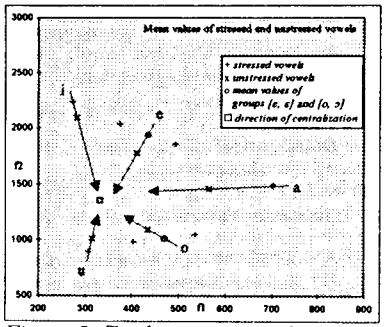


Figure 5. Tendency to centralization in unstressed vowels.

Many authors [7], [8], [9] have stated the importance of another kind of ideal point, named centroid and defined as "*...the grand mean of all measured formant frequency of the vowel system per speaker.*"[9:159]. The centroid is generally calculated either for only stressed or for only unstressed vowels, and gives an idea of a tendency that is internal to that vowel system. Our procedure, on the contrary, describes a tendency related to the different behaviour of the speakers when uttering a stressed vowel and an unstressed one. We calculated the coordinates of the centroid (average for all twenty speakers) of our stressed and unstressed vowels in order to compare them with the ideal point (from now on CT 'Centralization Tendency') calculated with our method. The results are:

centroid:		
stressed	f1 = 441	f2=1505
unstressed	f1 = 401	f2=1486
CT :	f1=332	f2=1350

CT seems to indicate a point in the f1/f2 plane quite different from the centroids. The ideal vowel corresponding to CT is higher and more velar (back) than the centroids.

Vowel duration

As expected, stressed vowels show a much longer duration (averagely about 97 ms) than final (62 ms) and non final unstressed (57 ms) ones (see Fig.6). It's very interesting to observe, on the other hand, that all groups of vowels show a very regular correlation between duration and openness. This correlation seems to be particularly relevant for stressed vowels, which range from a minimum of 84±29 ms for [i] and 86±29 ms for [u], to a maximum of 119±38 ms for [a].

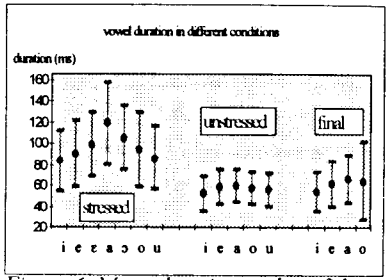


Figure 6. Mean duration and confidence interval of Italian vowels.

A comparison between the duration of stressed vowels in open syllables with vowels in closed syllables is shown in Figure 7. Unlike what is generally accepted in literature [2], [4], [10], [11] our data show that no significant difference exists between the two groups.

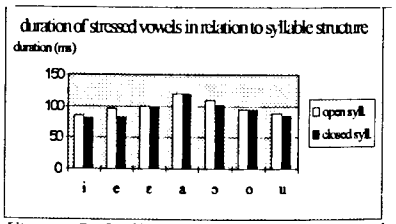


Figure 7. Differences of stressed vowel duration between open and closed sillables.

COMMENT

Comparing our stressed vowel data to the unstressed ones a clear tendency to centralization is observable. We have tried to express this tendency in terms of coordinates of a particular ideal point as shown in Fig. 5. As a consequence, our method seems to state the loss of the symmetry introduced with the concept of centroid.

Moreover, centralization is classically used in synchronic and diachronic phonetics for phenomena of vowel alternation or change resulting into a *schwa* [ə] (corresponding approximately to f1=500, f2=1500 Hz). In our case, on the contrary, we use the term centralization to indicate a tendency, i.e. a process of partial convergence towards an ideal point in the f1/f2 plane; which does not coincide with *schwa* (in this sense we use the term "point" instead of "vowel" in order to emphasize the difference).

The vowel [ə] has a well distinct nature and has also a pertinent phonological value in many languages. Presently Italian seems not to belong to the class of the languages having a [ə] in their vocalic system. Though, on the base of the results herein presented, a class of partially centralized vowels should be introduced for a more complete description of unstressed (non final and final) vocalic sounds.

In this paper many results seem to differ (more or less slightly) from the normally accepted description of Italian vowel system. It is our opinion that this is mainly due to the choice of using "spontaneous" speech materials instead of laboratory speech.

Many further investigations need to be carried out on our data, some of the further planned developments are:

-analyses of female data and description of the differences between the two groups;

-effects of coarticulation with adjacent consonants and of assimilation with vowels in adjacent syllables (some relations between these effects and the entity of the overlaps of vocalic areas are foreseen);

-investigations on the role of other measured acoustical parameters such as f_0, f3, maximum energy within the vowel;

-diatopic analyses of data, namely: a comparison between the vowel systems in different regional standards of Italian.

REFERENCES

[1] Ferrero, F. (1968), "Diagrammi di esistenza delle vocali italiane", *Alta Frequenza*, vol. 37, pp. 9-31.

[2] Fava E., Magno Caldognetto E. (1976), "Studio sperimentale delle caratteristiche elettroacustiche delle vocali toniche e atone in bisillabi italiani", in R. Simone et al. (eds), *Studi di fonetica e fonologia*, Roma, Bulzoni 1976, pp.35-79.

[3] Ferrero F., Magno Caldognetto E., Vagges K., Lavagnoli C. (1978), "Some acoustic characteristics of the Italian vowels", *Journal of Italian linguistics* 3, pp.87-96.

[4] Marotta G. (1984), "*Aspetti della struttura ritmico-temporale in italiano. Studi sulla durata vocalica*", Pisa, ETS.

[5] Albano Leoni F., Maturi P., (1994), "Didattica della fonetica italiana e parlato spontaneo", in Giacalone Ramat A., Vedovelli M., (eds) *Atti del XXVI Convegno nazionale SLI*, Roma, Bulzoni, pp.153-164.

[6] Albano Leoni F., Caputo M.R., Cerrato L., Cutugno F., Maturi P., Savy R., (1994), "Il vocalismo dell'italiano. Analisi di un campione televisivo", in Perrone B., (ed) *Atti del XXII Convegno nazionale AIA*, Lecce, pp.419-424.

[7] Liljencrants J., Lindblom B., "Numerical simulation of vowel quality systems: the role of perceptual contrast", *Language*, Vol.48, 1972, pp.839-862.

[8] Disner S.F., (1980), "Insights in vowel spacing: results of a language survey.", *UCLA Working papers in phonetics*, Vol.50, pp.70-92.

[9] Koopmans-van Beinum F.J., (1983), "Systematics in vowel systems", in van den Broecke M., van Heuven V., Zonneveld W., (eds) *Sound Structures*, FORIS Publications, Dordrecht, pp.159-171.

[10] Salza P.L., Sandri S. (1987), "Consonant-to-vowel durational effects in Italian", *CSELT Technical Report*, Vol. XV n.1, February, pp.61-66.

[11] Farnetani E., Kori S. (1984), "Effects of syllable and word structure on segmental duration in spoken Italian", *Quaderni del Centro di studio per le ricerche di fonetica*, 3 , pp.143-188.

A Microsoft Excel 5.0 version of our database may be accessed via anonynous ftp on Internet. The address of ftp server is: dsna1.na.infn.it. Go to the directory CIRASS and use binary mode to retrieve the file named vocali.xls.

Acknowledgements This work was realized with the help of all CIRASS members for measurements and data elaboration. Moreover we wish to thank Loredana Cerrato and Pietro Maturi that took part to all the preceeding phases of the project.

CONSONANTS AND VOWELS INFLUENCE ON PHONATION TYPES IN ISOLATED WORDS IN STANDARD CHINESE

A. Belotel-Grenié, M. Grenié
Université de Nice-Sophia Antipolis,URA CNRS 1235, Langues, Langage & Cognition
1361 Route des lucioles, F-06560 Valbonne-Sophia Antipolis France
e-mail : belotel@tana. unice. fr

ABSTRACT

The aim of this paper is to study the effects of consonants and vowels on phonation types in Standard Chinese. Tone 3 and tone 4 are often associated with the production of creaky voice. Our results show significant differencies in the type of vowel involved in the production of creaky voice. Creaky voice is observed not only for low vowels but also for high vowels. In term of initial consonants mode of articulation, significant effects have been found in term of duration of vowels produced with creaky voice.

INTRODUCTION

Standard Chinese has four tones, and each syllable has one of these specified lexically. According to the F0 contour on the vowel, the four tones are : high level (tone 1, hereafter T1), rising (tone 2, T2), low falling (tone 3, T3) and falling (tone 4, T4) tones. Generally the phonetic realisation of tones are described in term of F0 contour, amplitude and duration [5]. However, despite extensive litterature on Chinese tones, little data have been published on the effect of tones on segmental realisations. In several Chinese dialects changes in phonation types are involved in production of tones [2, 3, 9]. Although changes in phonation types are not phonologically distinctive in Standard Chinese, we have shown in previous study [1] that vowel /a/ at tone 3 and 4 is often produced with creaky voice. Significant differencies have been found between tones in the measure of relative energy of the fundamental and the largest harmonic in the first formant (F1-H1) at the beginning, the middle and the end of the vowel. The main purpose of the present study is to examine the role of initial consonant and subsequent vowel in production of creaky voice for isolated monosyllabic words. Our study addresses the following questions : Do all kind of vowels affected by creaky voice ? What is the effect of initial consonant on vowel phonation ? The following section presents the speech materials, the next section shows the distribution of vowels produced with creaky voice according to tones. We will then discuss the effect of vowels on creaky voice. Finally, we will describe the effect of initial consonants on phonation types.

SPEECH MATERIALS

The speech materials used in this study was collected in a sound proof chamber by researchers of the Chinese Academy of Social Sciences. Nine Beijing native speakers (seven males and two females) speaking Standard Chinese as their primary language, recorded all the Chinese monosyllables (1279 monosyllables including zero-initial syllables) in isolation using a DAT. They were students of Beijing University. All recorded speech samples were digitized with a Digidisign™ audio card on Macintosh™ with 16 bits quantization using a 10 kHz sampling rate. The tokens were analyzed, segmented and hand-labelled using Signalyze™. In this paper we present preliminary results for only one male speaker who uttered 634 isolated words. Duration and fundamental frequency maximum and minimum were measured for vowels.

CREAKY VOICE DISTRIBUTION

Each utterance was listened and have been visually inspected from a CRT display that simultaneously presented waveform, F0, amplitude, zero crossing and spectrogram. It seems that the speaker produces different phonation types (modal voice and creaky voice). Our analyses of his creaky voice replicate the findings published in [6, 7]. Figure 3 shows the oscillogram, the fundamental frequency curve, the amplitude curve, the zero crossing curve and the spectrogram of the syllable "xiao" produced at tone 3 with creaky voice. Creaky voice is characterised on oscillograms by irregularly spaced pulses, on spectrograms by eneven vibrations of vocal cords and by the presence of energy in the higher frequencies, on amplitude curves by a decrease of the amplitude and a greater shimmer. Table 1 presents the number of words produced with creaky voice with regard to the total number of words for each tone. It confirms our previous results [1] : creaky voice is never produced for tone 1. It also shows that creaky voice primarily affects tone 3 (45.8%) and secondly tone 4 (10.5%). The fact that none creaky voice was observed for tone 2 is not surprising because we have shown in [1] that changes in phonation seems to be a gradual speaker dependent phenomenon.

Table 1 : Number of words produced with creaky voice for each tone for one speaker with regard to the total number of words uttered by this speaker.

Tone1 cv/tot	Tone2 cv/tot	Tone3 cv/tot	Tone4 cv/tot	Total cv/tot
0/158	0/120	89/194	17/162	106/634
0%	0%	45.8%	10.5%	16.7%

VOWELS AND CREAKY VOICE

There are about thirty five finals in Standard Chinese. Five of these consist of a single vowel, the remaining thirty finals are combinations of medials, main vowels and endings. That is to say that a final may be composed as many as three elements : a medial that is a short vowel sound or a glide, the main vowel that is the principle carrier of the syllable and the ending that is a short vowel or a nasal consonant.

Presence of creaky voice

Analysis of creaky voice distribution shows that all kinds of vowels could be affected by a change in phonation type. Several occurrences of creaky voice have been found for /i/ utterances. In order to examine the link between phonation types and vowels, vowels were splitted into three classes according to the main part of the compound vowel : low vowels, high front vowels, high back vowels (figures 1 and 2). A Chi-square was computed on these data. For tone 3, highly significant differencies were found among the three classes ($p<0.0001$). The differencies between the three classes are also significant ($p<0.005$) for tone 4. It appears on one hand, that low vowels are more affected by creaky voice than others, and on the other hand, that high front vowels are more associated with creaky voice than high back ones. This pattern was observed for both tone 3 and 4.

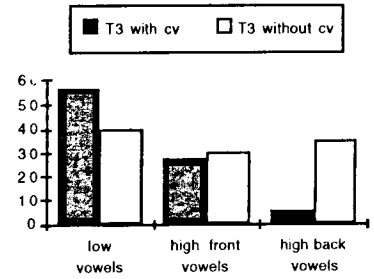


Figure 1 : Occurrences of vowels produced with and without creaky voice phonation at tone 3. The vowels are splitted into three categories (low vowels, high front vowels and high back vowels) according to their main vowel. Differencies between classes are highly significant ($p<0.001$).

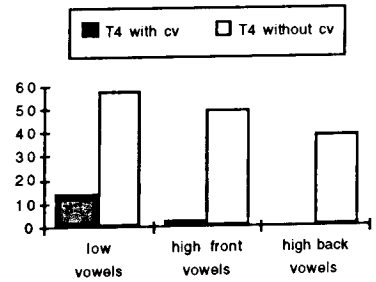


Figure 2 : Occurrences of vowels produced with creaky voice phonation and without creaky voice at tone 4. The vowels are splitted into three categories (low, high front and high back) according to their main vowel. Differencies between classes are significant ($p<0.01$).

Vowel duration and creaky voice

It has been known for many years that vowel duration in Standard Chinese depends on tones [6]. Table 2 presents the vowel mean duration according to tones. Our results are identical to those

observed in [6]. An analysis of variance shows that these variations in duration are highly significant ($p<0.0001$). The longer duration have been found for tone 3 and the shorter one for tone 4 . We observed the hierarchy T3>T2>T1>T4. It is therefore important to evaluate if creaky voice has an effect on vowel duration. An analysis of variance revealed that there is no significant effect of creaky voice on vowel duration.

Table 2 : Vowels mean duration (in ms) according to tones. The differencies are highly significant ($p<0.0001$).

mean	tone 1	tone 2	tone 3	tone 4
duration (in ms)	235.4	281.2	342.7	193.7

Table 3 : Vowels mean duration (in ms) with and without creaky voice for tone 3 and 4. Differencies are not significant.

	tone 3	tone 4
with cv	339.2	199.8
without cv	345.8	193.0

CONSONANTS EFFECTS ON VOWELS WITH CREAKY VOICE

There are twenty-one initial consonants in Standard Chinese. Unlike most European languages, Standard Chinese has no distinctively voiced consonants. There is a primary distinction between obstruent (stops, affricates and fricatives) which are all voiceless and sonorants (nasals, laterals and semivowels) which are all voiced. Stops and affricates falls into two contrasting series : aspirated one and unaspirated one. Considering the place of articulation there are five labial consonants, three alveolars, three dental sibilants, four retroflexes, three palatals, three velars. To complete analyses of our data, consonants have been splitted into categories according to their mode and place of articulation.

C.V. distribution and consonant mode and place of articulation

A Chi-square on creaky voice distribution according to mode of articulation does not reveal for tones 3 and 4 any significant effect.

A Chi-square was carried out on the data taking into account the place of articulation. The distribution of vowels with creaky voice and without creaky voice was not significantly different according to their place of articulation.

Effects of consonants articulation on duration of vowels with C. V.

An analysis of variance was run taking into account the mode and the place of articulation of consonants in order to determine whether they have an effect on vowel duration. The results showed a significant effect of the mode ($p<0.001$) on vowels with and without creaky voice. A close examination reveals that these differencies may be mainly due to the effect of aspirated vs unaspirated consonants ($p<0.0001$). After aspirated consonant the vowel is shorter than after unaspirated one. No significant effect of the place of articulation both for tones 3 and 4 have been found.

Effects of consonants articulation on the F0 of vowels with C. V.

Different analyses of variance were carried out to examine whether the articulation of initial consonant have an effect on the F0 values. The results showed that the mode of articulation of initial consonant have no significant effect on the F0 maximum and minimum values, that is to say that mode does not interfere on the F0 of vowels produced with and without creaky voice both for tones 3 and 4. However significant differencies ($p<0.005$) have been found between vowels with creaky voice and without creaky voice in term of F0 maximum. Vowels produced with creaky voice have a F0 maximum value lower than which are not creaky.

CONCLUSION

Changes in phonation types occur in Standard Chinese for all kinds of vowels. Low vowels are more produced with creaky voice than high ones and high front present more occurences of creaky voice than high back ones. Neither mode of articulation nor place of articulation influence phonation changes. Vowel duration shows a significant interaction between phonation type and initial consonant mode of articulation. The analysis of more data (several speakers and temporal measures of pitch) is needed to confirm these preliminary results.

ACKNOWLEDGEMENT

Authors are most grateful to all the members of the Phonetic Laboratory of Beijing Chinese Academy of Social Sciences, in perticular Prof. Lin Maocan, for the use of their speech recordings.

REFERENCES

[1] Belotel-Grenie A., Grenie M. (1994), Phonation types analysis in Standard Chinese. *Proceedings of ICSLP'94*, Yokohama, Japon, pp. 343-346.

[2] Cao J., Maddieson I. (1992), An exploration of phonation types in Wu dialects of Chinese. *Journal of Phonetics*, 20, pp. 77-92.

[3] Davison D. S. (1991), An acoustic study of so-called creaky voice in Tianjin Mandarin. *UCLA, Working Papers in Phonetics*, 78, pp. 50-57.

[4] Huffman M. K. (1985), Measures of phonation type in Hmong. *UCLA, Working Papers in Phonetics*, 61, pp. 1-25.

[5] Howie J. M. (1976), *Acoustical studies of mandarin vowels and tones*. Cambridge University Press.

[6] Kirk P. L., Ladegoged P., Ladefoged J. (1984), Using a spectrograph for measures of phonation types in a natural language, *UCLA Working Papers in Phonetics*, 59, pp. 102-113.

[7] Ladegoged P., Maddieson I., Jackson M. (1988), Investigating Phonation types in different languages, in *Vocal folds physiology, voice production, mecanisms and functions*, ed. Fujimura O., Ravers Press.

[8] Maddieson I., Ladefoged P. (1985), "Tense" and "Lax" in four minority languages of China. *Journal of Phonetics*, 13, pp. 433-454.

[9] Svantesson J. O. (1990), Initial consonants and phonation types in Shanghai. *Phonetic Experimental Research at the Institute of Linguistics*, University of Stockholm, XIII, pp. 1-4.

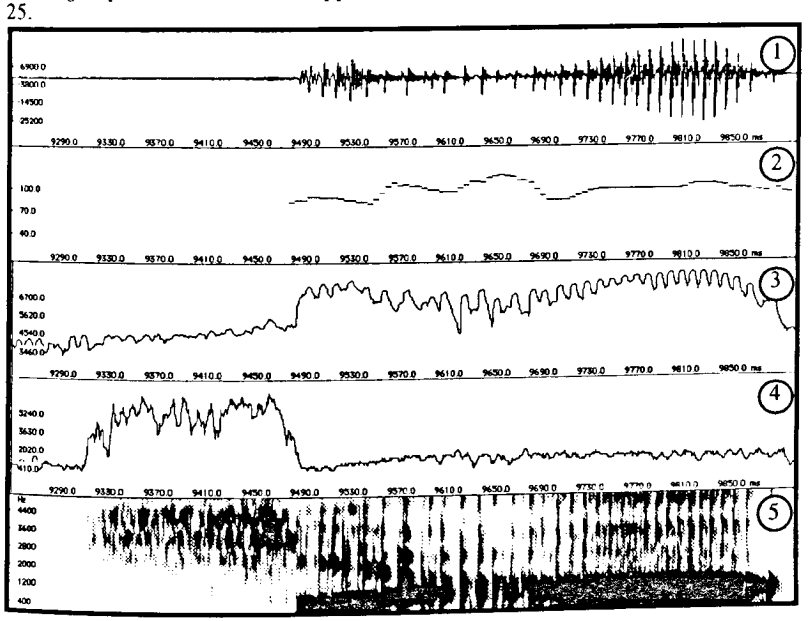


Figure 3 : Acoustic analysis of the syllable "xiao" (in pinyin) /cja^{u}/ produced at tone 3. Oscillogram in (1), fundamental frequency in (2), amplitude curve in (3), zero crossing curve in (4) and large band spectrogram in (5). The presence of creaky voice can be seen at the middle of the word. F0 detection failed to detect the real value of pitch during the production of creaky voice. Amplitude curve shows a greater shimmer during the creaky phase.

ACOUSTIC CHARACTERISTICS OF GREEK VOWELS UNDER DIFFERENT PROSODIC CONDITIONS

Antonis Botinis, Marios Fourakis** and Maria Katsaiti*+*
**Phonetics Laboratory, Dept. of Linguistics, University of Athens, Greece*
***Dept. of Speech and Hearing Science, Ohio State University, USA*
+Names in alphabetic order

ABSTRACT

Five male native speakers produced five repetitions of words containing the five vowels of Greek when stressed, unstressed, or in focus and at two tempi of speech, normal and fast. Measurements were made of the first three formants, duration, F0, and overall amplitude. Main results: 1. The vowel space expanded in focus and contracted in the absence of stress. 2. F0 was heightened for vowels in focus, stress, and fast tempo. 3. Focus did not affect durations but tempo and stress did.

INTRODUCTION

The present paper reports an acoustic analysis of the Greek vowels. The analysis includes spectral, durational, fundamental frequency (F0), and amplitude correlates under different prosodic conditions of tempo, stress, and focus. Aspects of the acoustics of the Greek vowels have been reported separately by Fourakis [1], Botinis [2], and Jongman, Fourakis, and Sereno [3]. Fourakis [1] studied the effects of tempo and stress on duration and reported a similar effect (25%) of these two conditions on duration. The effect of the stress condition on duration was replicated by Botinis [2]. Botinis [2] studied the effects of stress and focus on the distribution of prosodic correlates and reported duration combined with intensity and F0 as the main acoustic correlates of stress and focus conditions respectively. The spectral correlates reported by Jongman et al [3] showed that the Greek vowels, when bearing lexical stress, are well separated in the acoustic space, allowing for maximal contrast between vowel categories. However, Jongman et al [3] did not examine formant characteristics under different conditions of tempo and stress. In this experiment, the effects of these variables on Greek vowels are analysed in a single experiment combining all conditions into one design.

EXPERIMENTAL METHOD

Speakers. The speakers were five Greek male students, with some knowledge of other languages (mostly English), who were recruited at Athens University. They spoke standard (Athenian) Greek and were between 20 and 23 years old.

Speech material. The test words were lexical stress minimal pairs. When a minimal pair could not be found, an extra word of similar structure was used to control for the difference. All words started with voiceless [p] followed by the target vowel and one or two voiceless obstruents (Table 1 below).

Table 1. Test words of minimal stress pairs and control words.

Stressed	Unstressed	Control
'pisa	pi'sta	'pista
(tar)	(loyal)	(track)
'pese	pe'ta	'peta
(fall)	(throw)	(fly)
'pasa	pa'sa	
(pass)	(pasha)	
'posa	po'sa	
(how)	(amounts)	
'pusi	pu'stça	'pusti
(fog)	(shabby trick)	(gay)

There were two conditions of elicitation. In one the subjects read lists containing the target words in the carrier sentence: [to 'sinθima "target word" tus a'resi po'li] 'they like the password "target word" a lot'. In the other condition the subjects were asked to respond to the question: [pço 'sinθima tus a'resi po'li] 'Which password do they like a lot?'. In the response, which was the same as above, the target word appeared in focus position. Only the words with stressed first syllables were used in this condition. The lists for each condition contained five repetitions of each target word and were read at a normal and a fast tempo with different randomisation of the sentences for each speaker and for each tempo of speech.

Measurements. All measurements were made using the Kay Elemetrics CSL hardware/software combination at Athens University Phonetics Laboratory. Utterances were digitised at 10KHz sampling rate with 16 bit resolution and measurements were made as follows:

1. An FFT was done using CSL's default settings at the middle of the vowel duration and the first three peaks in the resulting spectrum were measured. In addition, whenever necessary, the FFT was supplanted by LPC analysis.

2. Vowel duration was measured from the waveform from the first glottal pulse after the release burst of the initial stop to the cessation of all discernible voicing before the following obstruent.

3. The duration of three glottal pulses in the middle of the vowel was measured, and the period and the F0 were computed. In the case of very short vowels (unstressed, fast tempo) with less than three glottal pulses, all available pulses were used. Some productions did not include any appreciable voiced interval and were excluded from the analysis.

4. A measure of the overall amplitude of the target vowel portion was computed in dB RMS.

RESULTS

1. Spectral characteristics. Figures 1a and 1b show an F1 by F2 acoustic space in which the positions of the vowels are plotted by the mean frequencies of their formants at the normal and fast tempo when stressed, unstressed, or in focus. Two acoustic effects are evident. First, there is a compression of the acoustic space under the unstressed condition in terms of vowel scattering on the F1 and F2 frequency axis. This effect is evident under both normal (Fig. 1a) and fast tempo conditions (Fig. 1b). Second, there is an acoustic raising under the unstressed condition in terms of an F1 frequency decrease. This is evident for all vowels and both tempi except for the vowel [u] at the normal tempo (Fig. 1a). An additional F1 decrease under 300 Hz is caused under the fast tempo condition for the high vowels [i] and [u] (Fig. 1b).

In order to evaluate the global effect of tempo, stress, and focus on the vowel space as a whole, the area of the space expressed in Hz-squared was computed. This technique has also been employed by Fourakis [4] and Bradlow [5]. Table 2 below shows the results expressed as ratios of the vowel space in each condition to the vowel space in the normal-stressed condition.

Table 2. Ratios of vowel space under different prosodic conditions.

Condition	Ratios
normal-stressed	1.00
normal-unstressed	0.73
fast-stressed	1.07
fast-unstressed	0.87
normal-focus	1.29
fast-focus	1.09

FIGURES 1a AND 1b (FORMANTS)

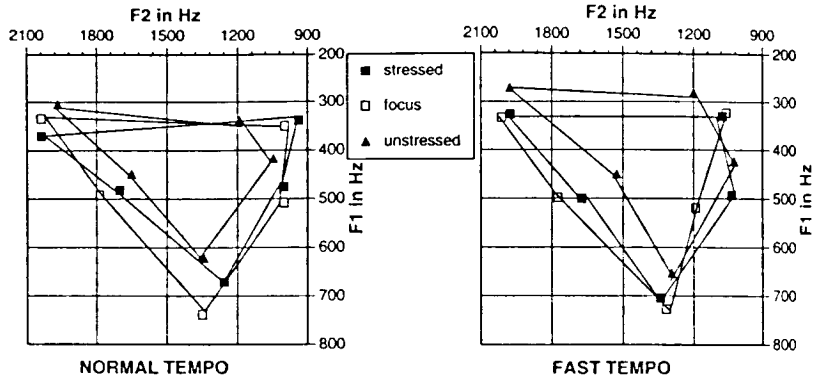


FIGURE 2 (DURATIONS)

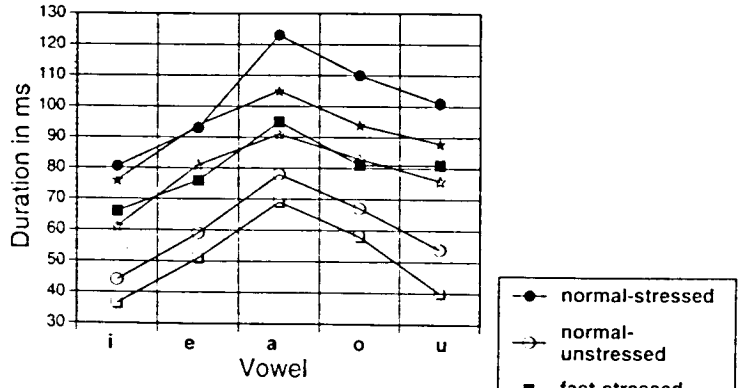
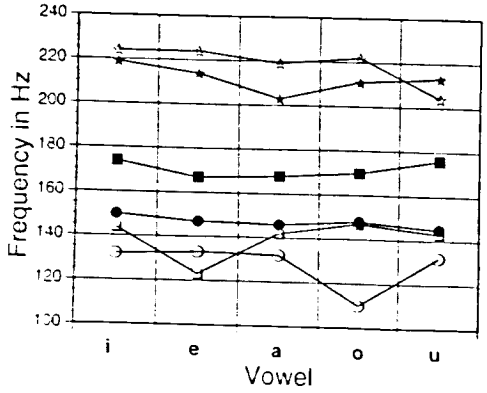


FIGURE 3 (F0)



The ratio values indicate two main effects. First, the vowel space is contracted when the vowels are unstressed regardless of tempo. Second, the vowel space is expanded when the vowels are in focus regardless of tempo.

2. Durations. Figure 2 shows the mean durations of each vowel for each condition of tempo, stress, and focus. Unstressed vowels are 40% shorter than stressed vowels at each tempo. Vowels spoken at the fast tempo are on the average 30% shorter than at the slow tempo when stressed, and 15% shorter when unstressed. Vowels in focus display a more complex pattern. At the normal tempo there is no difference between stressed and in-focus front vowels, but there is a considerable difference in the back vowels. Stressed vowels are longer than vowels in focus. At the fast tempo, there is no difference for any of the vowels. In summary, normal tempo vowels are longer than fast tempo vowels, stressed vowels are longer than unstressed vowels, and vowels in focus are as long as stressed vowels except for back vowels at normal tempo.

3. Fundamental frequency. Figure 3 shows vowel F0 for each vowel in each condition. Three effects are clear. First, vowels in focus have much higher F0 than in any other condition, regardless of tempo. Second, vowels at fast tempo have higher F0 than vowels at normal tempo and this is most regular under the stressed condition. Third, stressed vowels have higher F0 than unstressed vowels, regardless of tempo. No other condition has greater effect on F0 than the focus condition. This is a strong evidence that F0 is the main acoustic correlate of focus in Greek.

4. Amplitude. The results of vowel amplitude expressed in dB RMS show a regular distribution of higher amplitude for stressed vowels (a detailed acoustic analysis of the Greek vowels is forthcoming [6]).

CONCLUSIONS

The results show that tempo, stress and focus may compress, expand, or raise the acoustic space. These acoustic variations do not however reach the phonetic level of vowel distinctions at Athenian Greek. The effect of focus and the effect of tempo on formant structure have not been reported, to our knowledge, in acoustic literature on other languages. Neither has the effect of tempo on F0. On the other hand, the effect of stress and tempo on duration, the effect of stress on amplitude, and the effect of focus on F0 in Greek have been corroborated by the present investigation.

REFERENCES

[1] Fourakis, M. (1986), "An acoustic study of the effects of tempo and stress on segmental intervals in Modern Greek", *Phonetica*, vol. 43, pp. 172-188.

[2] Botinis, A. (1989), *Stress and prosodic structure in Greek*, Lund: Lund University Press.

[3] Jongman, A., Fourakis, M. and Sereno, JA. (1989), "The acoustic vowel space of Modern Greek and German", *Language and Speech* 32, pp. 221-248.

[4] Fourakis, M. (1991), "Tempo, stress and vowel reduction in American English", *J. Acoust. Soc. Am.* 90, pp. 1816-1827.

[5] Bradlow, A. (1995), "A comparative acoustic study of English and Spanish vowels", *J. Acoust. Soc. Am.* 97, pp. 1916-1924.

[6] Fourakis, M., Botinis, A. and Katsaiti, M. (forthc.), "Acoustic correlates of the Greek vowels".

LABORATORY VS SPONTANEOUS SPEECH A CASE STUDY OF SONORANTS

M. Chafcouloff and A. Marchal
URA 261, CNRS Parole et Langage
Institut de Phonétique, Aix-en-Provence, France

ABSTRACT

A study was conducted to analyze the acoustical characteristics that distinguish French sonorants in laboratory vs. spontaneous speech conditions. A dialogue was set up to 'elicit' answers from a speaker who uttered lexical words with /jwlr/ in initial and final position in a vocalic context /i,a,u/. Results showed that steady-state duration is significantly shortened in spontaneous speech, whereas transition duration is less affected by changes in speaking rate and stress. No significant differences in F2 values were found across speaking styles, which means that the concept of reduction does not apply to the production of sonorants in French. Results are then discussed in relation to the target undershoot model.

INTRODUCTION

During the past decades, phonetic research has mostly privileged the use of a peculiar kind of speech, namely 'Laboratory speech', i.e. nonsense words or lexical words uttered in isolation or embedded in a carrier sentence, to study the acoustical characteristics of speech sounds. Yet, deceptive results in text-to-speech synthesis and speech recognition systems have led researchers to conclude that the cues extracted from such speech signals were insufficient carriers of 'real' speech. In other terms, and it was a message clearly expressed during the last ICPhS in Aix-en-Provence, it was urgent to move away from laboratory speech to study a more natural speech.

In the study of the acoustic/phonetic characteristics that distinguish laboratory speech from spontaneous speech, much work has concerned vowel reduction; contrary to the results of some previous studies, it was found that short durations due to a faster speaking rate did not necessarily result in formant undershoot, notably in Dutch [10].

So far, most quantitative data about the acoustical characteristics of the sonorants /jwlr/ has been obtained from the analysis of laboratory speech samples [9,5,2]; for a detailed review of acoustic and perceptual work, see [6].The study of the effects of suprasegmental factors as speaking rate and stress, has led to controversial results.Whereas Klatt [8] reported noticeable formant undershoot in English, Chafcouloff [3] found no significant differences in French. As both studies were concerned with the analysis of laboratory speech items, it is of interest to inquire how these sounds behave acoustically in different conditions of speech.

Actually, several questions may be asked:
-Does a change in speaking style drastically affect the formant structure of sonorants in French?
-Does the concept of reduction apply to the production of these sounds?
-Are there any recurrent acoustical characteristics which may allow to distinguish laboratory speech from spontaneous speech?

METHOD AND SPEECH MATERIAL

In order to build up a solid base of comparable data, a controlled elicitation method of spontaneous speech was used. A question-answer dialogue was set up.The recording took place in an anechoic room, where a speaker of southern French was seated in front of the investigator. The role of the latter was to keep the conversation fluent, and to ask questions until the speaker uttered the 'expected' word. Secondly, the same

thirty-one lexical words previouly uttered in the dialogue, were read twice by the speaker. The sonorants were found in the initial and final position, e.g.yak vs. paille, in a vocalic context /i,a,u/ and varying stress position, e.g. 'loup vs. lou'bard.

A listening experiment aimed at assessing the naturalness of the utterances was organized. Two speakers met the requirements. Their speech was judged as typical of a spontaneous speech situation, but the third speaker failed the test. Consequently, the results reported here pertain to the data of two speakers only.

A prosodic analysis was conducted for displaying the Fo configurations of the sentences uttered in spontaneous speech. Three main intonative patterns were used:
1. When the speaker was somewhat wavering, his answer was a question for seeking confirmation. In this case, the word lies at the end of the sentence, and a rising intonative pattern is used (62%)
2. When the speaker enumerated several words which might correspond to the answer, the intonative pattern was usually flat (18%).
3. When the speaker was utterly confident of giving the right answer, a declarative falling pattern was used (20%).

The acoustic analysis was based on the use of an editing program. The utterances were digitized using a 10 Khz sampling rate with 12-bit resolution. Speech signals were pre-emphasized to compensate for weak spectral energy of sonorants at high frequencies. Wideband spectrograms were made and formant frequencies were calculated through FFT and LPC analysis.

RESULTS

Temporal characteristics

Measurements made from oscillographic tracings and spectrograms revealed that lexical words were around 15% shorter in spontaneous speech than in laboratory speech. Average word duration pooled over the two speakers was 292ms for spontaneous speech vs.347ms for laboratory speech. This demonstrated that a faster speaking rate was generally used in spontaneous speech (average 6.6syl./sec.) compared to laboratory speech (5.2 syl./sec.).

Figures 1 & 2 illustrate average steady-state and transition duration values for /jwlr/ across speaking styles. One notices that steady-state portions are significantly shorter ($p < 0.1$) in spontaneous speech than in laboratory speech, and that a 2:1 duration ratio is most often observed. This is especially true concerning the /l/-sound which is characterized by the longest steady-states (>100ms) and the shortest transitions (<30ms). However, if the /l/'s duration is relatively constant across speakers, attention must be drawn on the fact that there is a great deal of intra and inter speaker variation for initial /jwr/ which is not reflected on the figure. Steady-state duration of /j/ and /w/ measured from other speech items in spontaneous speech may be as short as 20ms which merely corresponds to 2 or 3 glottal pulses along the time axis; conversely, it may be as long as 80ms when the word is uttered with a strong emphatic stress. Likewise, the steady-state of /r/ varies as a function of the relative duration of a schwa-vowel initial segment; this variety is often found in the allophones of /r/ in southern French [4].

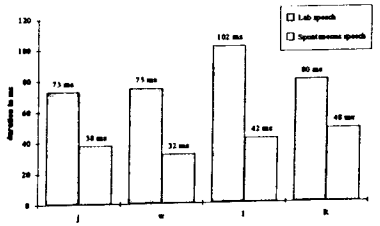


Figure 1. Steady-state durations. Mean duration is pooled over 2 syllable positions, 2 speakers and 3 vowel contexts.

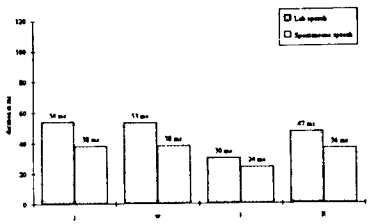


Figure 2. Transition durations. Mean duration is pooled over 2 syllable positions, 2 speakers and 3 vowel contexts.

As far as duration of transitions is concerned, it should be noticed that the transitions of /l/ are shorter than that of /jwr/ in both speaking styles (differences are significant at $p < 0.5$).Concerning the ratio between steady-state and transition, whereas the steady-state portion of /l/ is usually twice longer than the transition, it turns out that the transitions are of approximately equal duration for /jwr/ in spontaneous speech, which is not the case in laboratory speech.While the lateral's spectrum is essentially static, the glides' is mainly dynamic. Thus, it appears that the transition is affected to a lesser degree than the steady-state by changes in speaking rate. In relation with this, the correlation coefficients are small for /l/ ($r(8) = 0.567$) and high for /jwr/ ($r(8) = 0.799$, $p < 0.1$).

Spectral characteristics

Differences in terms of vocalic space along a F1/F2 dimensional plane for /jwl/ are illustrated on Figure 3. As the /r/-sound was mostly produced as a fricative allophone with a predominant noise source and no clear-cut formant structure, the results pertaining to the /r/-sound have not been included. As shown by the closeness of the points on the chart, it can be observed that these are clustered in three relatively compact areas, and that the acoustic distance separating the white from the black squares is mostly short. Thus, it seems that neither stress, nor speaking rate exert a decisive influence on the formant frequencies, as no statistically significant differences were found between F2 values for sonorants uttered in different speaking styles.

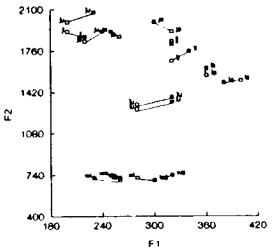


Figure 3. Acoustic distance between /jwrl/ uttered in laboratory (□) vs. spontaneous speech (■).

In these conditions, is it to say that there are no changes at all concerning the production of sonorants? As a matter of fact, a closer examination reveals subtle modifications in the acoustic spectrum. As mentionned above, /j/ and /w/ may be uttered in spontaneous speech as very brief segments. From our data, it appears that /j/ may be produced as a voiced obstruent especially in the /i/ context which is not the case in laboratory speech. Moreover, /j/ is characterized by asynchronous movements in the F-pattern especially at the third formant level. This observation brings credit to earlier remarks from other authors who have noticed the complex temporal evolution of formants. This is true for /j/ [1], for /l/ [7], and for /w/ [2]. Contrary to laboratory speech, the release of the /l/-sound is often characterized in spontaneous speech by a transient click in the upper part of the spectrum. This remark is consistent with the earlier observations of Dalston [5] who suggested that a noise transient associated to a rapid release of the apex away from the alveolar ridge might be an important cue for the identification of this sound. In addition, we have found several cases of formant continuity/discontinuity in the /l/'s spectrum both as a function of position and speaking style. Lastly, as the /r/-sound is most sensitive to contextual

effects [4], it is evident that any changes in speaking styles should be followed by subsequent changes in its acoustical structure. Observations have revealed changes in terms of predominance of an harmonic spectrum vs. a noise spectrum as a function of coarticulatory effects. As the /r/'s duration is shortened, the number of flaps across its stationary portion may be similarly affected. Finally, a retroflex allophone of /r/ was found in spontaneous speech, especially in a back-vowel context.

CONCLUSION

In the state of research, it must be acknowledged that the results presented here are limited in that our data pertain to a single utterance of thirty-one speech items uttered spontaneously by two speakers. Nevertheless, preliminary results indicate that the temporal characteristics of sonorants undergo changes as a function of different conditions of speaking rate and stress. However, these durational changes do not result in any systematic differences in formant frequencies especially at the F2 level. The fact that the sonorants' acoustic targets remain essentially unchanged, implies that the concept of reduction does not apply to the production of sonorants in French. This statement is not necessarily at variance with Klatt's findings who reports significant neutralization of formant target cues for /wjlrh/ in English [8]. Because of the basic tense-lax opposition between the two languages, one should expect that English sonorants tend to be more reduced than their French counterparts.

Moreover, as no frequency differences were found, despite the fact that that segmental duration was generally shorter in spontaneous speech, we may conclude that our results do not support the target undershooot model and its refined versions. Instead, they agree with the results of Van Son and Pols [10] who found no measurable relation between vowel duration and F2 frequency values in normal and fast speaking conditions. Thus, it seems that this model may not apply to all speech sounds across languages and also may not be valid for all speaking styles, especially in spontaneous speech.

REFERENCES

[1] Carlson,R. and Nord, L. (1990) "Analysis and Synthesis of Swedish sonorants-Part 2", Phonum (1), 70-73, University of Umea

[2] Chafcouloff, M. (1980) "Les caractéristiques acoustiques de /jwlr/ en français", Travaux de Phonétique de l' Institut de Phonétique d'Aix, 7, 7-56

[3] Chafcouloff, M. (1982) "L' effet du débit de parole sur les caractéristiques de /jwlr/", Travaux de Phonétique de l' Institut de Phonétique d'Aix, 8, 163-187.

[4] Chafcouloff, M. (1986) "Quelques variantes de /r/ en français méridional" Travaux de Phonétique de l' Institut de Phonétique d'Aix, 10, 101-120.

[5] Dalston, R.M. (1975) "Acoustic characteristics of english /wrl/ spoken correctly by young children", Journal of the Acoustical Society of America 57, 462-469.

[6] Espy-Wilson,C. (1992) "Acoustic measures for linguistic features distinguishing the semivowels /jwlr/ in American English", Journal of the Acoustical Society of America 92 (2), 736-757

[7] Fant, G. (1960) "Acoustic Theory of Speech Production", Mouton, the Hague.

[8] Klatt, D.H. (1974) "Acoustics characteristics of /jwlrh/", Journal of the Acoustical Society of America, 55 Suppl.1, Paper G11, p. 397.

[9] Lehiste, Ilse (1964) "Acoustic characteristics of selected English consonants", International Journal of American Linguistics 29, 1-197.

[10] Van Son, R.J. and Pols, L.C. (1992) "Formant movements of Dutch vowels in a text read at normal and fast rate", Journal of the Acoustical Society of America 92 (1), 121-127.

TOWARDS AN ACOUSTIC DESCRIPTION OF BRAZILIAN PORTUGUESE NASAL VOWELS

Elizabeth Maria Gigliotti de Sousa
Laboratório de Fonética Acústica e Psicolinguística Experimental
Universidade Estadual de Campinas

ABSTRACT

This study investigates the acoustic properties of nasal vowels in stressed syllables in Brazilian Portuguese (BP) according to experiments conducted at the State University of Campinas (UNICAMP). We concentrate here on the acoustic features of BP nasal vowels as opposed to BP oral vowels.

Data analysis concerns frequency (Hz) of the main vocalic formants as well as duration (ms).

INTRODUCTION

Nasality constitutes an important cue to the distinction of both consonants and vowels in Brazilian Portuguese (BP). The BP vocalic system in stressed syllables comprises seven distinctive oral vowels ([a, e, ɛ, i, ɔ, o, u]) and five distinctive nasal vowels ([ɜ̃, ẽ, ĩ, õ, ũ]). All those vowels occur in minimal pairs v x ṽ, such as [kata] x [kɜ̃ta] ('cata' x 'canta' = "he/she picks s.t. up" x "he/she sings") or [pita] x [pĩta] ('pita' x 'pinta' = "he/she smokes a pipe" x "dot"). Those are surface phonetic distinctions; there is actually a controversy as to the phonological status of nasal vowels in BP. This controversy, however, does not directly involve the acoustical descriptive framework presented in this paper.

Portuguese has non-distinctive nasal vowels as well, as in [sẽɲa] ('senha' - "password") and [sĩma] ('cima' - "over"). In these cases we may say that vowel nasality derives of anticipatory lowering of the velum to produce the following nasal consonant. Such an anticipation does not explain, however, the behavior of Brazilian phonetically distinctive nasal vowels. These vowels may appear both in nasal and non-nasal environments, what means they are not dependent on a following nasal consonant to be nasalized by BP speakers.

ACOUSTIC DESCRIPTIONS OF VOWEL NASALITY

Phoneticians have reported nasality in vowels as a rather difficult descriptive problem. The coupling of the oral and the nasal tracts involves a wide range of acoustic effects, depending on voice quality, oral and nasal tract volume and shape, degrees of coupling, etc. The nasal cavity introduces extra formants (poles) as well as anti-formants (zeros) in the acoustic output of vowel articulation. The extra poles and zeros change considerably the spectra of nasalized vowels when compared to similar oral vowels.

Some of the major acoustic changes, as reported by House & Stevens [1], Pickett [2], Curtis [3] and Hawkins & Stevens [4], are:

a) an extra pole-zero pair in the vicinity of F1 that interferes with this formant; b) general formant damping; c) weak formants and anti-formants surrounding F2 and F3.

Brazilian nasal vowels have traditionally been described in auditory terms. Early attempts of acoustic description of these sounds (Cagliari [5]) were thwarted by the lack of proper hardware and software tools that could enable more precise descriptions.

THE EXPERIMENT

Our first step towards the description of nasality in BP consisted of a series of preliminary studies featuring both nasal vowels and nasal consonants in BP words in different phrasal contexts. On the basis of those early experimental findings we built a more controlled experiment, this time featuring only distinctive nasal vowels.

The experiment consisted of [pV] monosyllables inserted in the carrier sentence "Digo ____ pra ele" ("I say ___ to him"). Each [pV] monosyllable consisted of either an oral or a nasal BP vowel. The resulting sentences were uttered three times by four BP male speakers coming from different parts of the country. By means of a Kay Elemetrics Sonagraph DSP - 5.500 we analyzed the total set of 84 oral vowels and 60 nasal vowels (three sets of oral/nasal vowels per speaker).

Our analysis featured: a) frequency (Hz) of F1, F2, F3, F4 and of the first nasal formant (Fn1); b) intensity (dB) of these formants; c) duration (ms): - of the monosyllables; - of vowels. For nasal vowels, we also measured the duration of the remarkable spectral changes observed in the end of most of the utterances. Such spectral changes were called "nasal murmurs".

We managed to compare formant intensity measures through a normalization procedure that set the F1 intensity value of each analyzed vowel as the reference to its other formant values.

Thus, if the [a] F1 intensity value were (-29 dB) and [a] F2 = -33 dB / [a] F3 = -44 dB, we would have:

Δ RI F2 = IF2 - IF1 = -33 - (-29)
Δ RI F2 = -4 dB

Δ RI F3 = IF3 - IF1 = - 44 - (-29)
Δ RI F3 = -15 dB
(I=intensity; RI=relative intensity)

Formant frequency analysis featured mainly 300 Hz and 150 Hz filters (sporadically involving 59 Hz and 117 Hz filters as well).

[pV] coarticulation was isolated whenever possible and non-included in vowel duration values.

After instrumental analysis we conducted a series of Student t-tests and variance analyses considering nasality, vowel quality and observation, in an attempt to extract a possible vowel nasality pattern. Those tests were conducted on the data for: a) F1 frequency; b) F2 frequency; c) vowel duration with and without the nasal murmur.

EXPERIMENTAL RESULTS

Formant frequency and intensity analyses

Although the BP oral subsystem has seven vowels and the nasal subsystem has only five, formant frequency and intensity analyses indicate that the relations among the vowels within each subsystem remain stable in spite of nasality. There is considerable difference, however, when similar vowels of each subsystem are compared in pairs, as in [po] x [põ], [pɔ] x [põ] and [pa] x [pɜ̃].

Variance analysis over vowel formant frequencies showed a dependency of nasality on vowel quality. Nasalization features change considerably according to the vowel with which they occur, thus rendering a consistent vowel-independent nasality pattern very hard to be achieved.

Our four speakers uttered the vowels [ẽ] and [õ] with varying degrees of opening; these vowels were also diphthongized in approximately 60 % of the utterances into [ẽʲ] and [õʷ].

Nasal vowels' formant intensity analysis showed a more leveled pattern than oral vowels', thus confirming the acoustic damping of nasal vowel formants that has been reported concerning other languages.

We found out that BP nasal vowels have a prominent nasal formant near F1; this formant accounts for the high acoustic energy level we found in the spectra of nasal vowels at low frequencies (-400 Hz). This first nasal formant (Fn1) displays constant acoustic energy level throughout the vowel and, in 87 % of the cases, its intensity was either similar to or higher than F1's. Minor nasal formants with low intensity appear in high frequencies, mainly between F2-F3 and F3-F4.

Duration Analysis

Duration analysis of BP nasal vowels showed the most interesting results.

A) Measures showed that distinctive BP nasal vowels are longer than their oral counterparts. Likewise nasal monosyllables are longer than oral ones.

B) Three distinct realization phases were identifiable in BP nasal vowels, namely: a) an oral release; b) a "nasalized phase" in which oral and nasal resonances are present; c) a "nasal murmur" phase in which nasal resonances prevail (see next page fig. 1).

C) The vowels [ĩ] and [ũ] displayed a longer nasal murmur phase than [ẽ] and [õ] (see fig. 2). The nasal murmur phase in [ɐ̃] was also considerably long, although not present in all the utterances.

Considering duration data presented in figure 2, we could assume that the greater duration of nasal vowels in relation to their oral counterparts could be credited to a greater duration of the nasal murmur phase.

One of the speakers, however, (a mid-western BP speaker) showed no final changes in the spectra of four nasal vowels (out of fifteen); his data also presented consistently shorter nasal murmurs when compared to other speakers'.

On these grounds we may assume that presence and/or duration of the nasal murmur phase may depend heavily on dialectal factors.

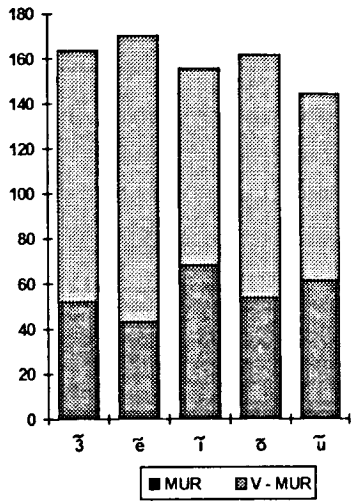


Figure 2. Nasal vowels' duration: full vowel duration, nasal murmur duration.

CONCLUSION

More studies are needed to account for all the variety of environments in which distinctive BP nasal vowels can occur (e.g., two/three syllable words, when followed by fricatives and stops, etc). Studies on dialectal and individual variation should also be carried out in order to properly portray nasalization in BP.

This study may be regarded as an initial attempt on the way to learning, within an Acoustic Phonetics framework, how nasality in Brazilian Portuguese effectively works.

ACKNOWLEDGMENT

This work was supported by CNPq grant number 50.0400/90 and FAPESP grant number 93/0565-2 to Eleonora Cavalcante Albano and by a CNPq/UNICAMP master's program scholarship to the author.

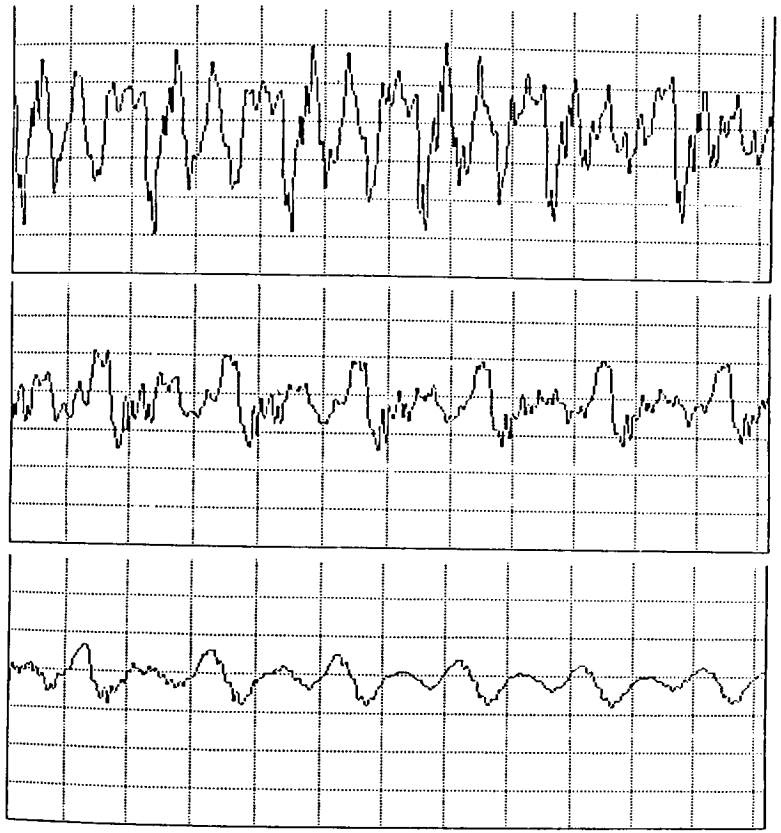


Figure 1. Waveform for vowel [ɐ̃]: phase 1 - oral release (upper), phase 2 - nasalized vowel (middle) and phase 3 - nasal murmur (lower).

REFERENCES

[1] House, A. S. & Stevens, K. N. (1956), "Analog studies of the nasalization of vowels", *Journal of Speech and Hearing Disorders*, 21, pp.218-232.

[2] Pickett, J. M. (1991), "Consonants: nasal, stop and fricative manners of articulation" in: *Readings in Clinical Spectrography of Speech*, San Diego, Singular Publishing Group & Kay Elemetrics Corp., pp.113-123.

[3] Curtis, J.F. (1970), "The acoustics of nasalized speech", *Cleft Palate Journal*, 7, pp.380-396.

[4] Hawkins, S. & Stevens, K.N. (1985), "Acoustic and perceptual correlates of the non-nasal/nasal distinction for vowels", *Journal of the Acoustical Society of America*, vol. 77 (4), pp.1560-1574.

[5] Cagliari, L.C. (1977), *An experimental study of nasality with particular reference to Brazilian Portuguese*, Ph.D. thesis (unpublished), University of Edinburgh.

[6] Sousa, E.M.G. de (1994), *Para a caracterização fonético-acústica da nasalidade no Português do Brasil*, M.A. dissertation, State University of Campinas.

PHONETIC ANALYSIS OF VOWEL SEGMENTS IN THE PHONDAT DATABASE OF SPOKEN GERMAN

S. Heid, M.-B. Wesenick, Chr. Draxler
IPSK – Institut für Phonetik und Sprachliche Kommunikation, Munich, Germany
e-mail: {heid, draxler, wesenick}@phonetik.uni-muenchen.de
http://www.phonetik.uni-muenchen.de

ABSTRACT

This paper describes some characteristics of the PhonDat database of spoken German [9] and its use for empirical studies in phonetic research. As an example results of vowel duration and formant measurement are presented.

INTRODUCTION

The aim of this paper is to show how symbolic data related to speech signals can be made available in a well-structured way and how this information can be used to determine and extract the relevant signal fragments for an acoustic analysis. Presently methods for the investigation of very large speech corpora are being developed. As an example, we retrieve from the symbolic database information on vowel position and compute the duration of vowel classes in selected consonant contexts. We then use the position information to apply a semi-automatic formant extraction program to the signal fragments. Duration measurements of 9950 vowels with specified contexts and formants of 10629 vowels are presented.

The purpose of our investigations currently is to establish a basis for the empirical study of German phonetics and phonology.

SPEECH DATA

The PhonDat Database (PhDB) consists of two main corpora of which only one - the PhonDat II train enquiry corpus - was investigated for the present studies. It contains data of 16 speakers with 64 read sentences each from the domain of train enquiries. The speech signals of the utterances have been segmented manually using a broad phonemic transcription (SAMPA) relative to the given citation form. The PhDB strictly adheres to the Computer Representation of Individual Languages (CRIL) guidelines agreed upon at the IPA Kiel 89 convention. Data is represented on three different symbolic levels: orthography, citation form, and phonetic transcription with time marks. The PhDB is implemented in Prolog [2], using the persistent Prolog environment Eclipse [3]. Access to the data is possible via the symbolic data levels; the result of a query is a reference to a signal fragment, or again symbolic data. Most database queries can be formulated using the query toolbox with little Prolog knowledge (except for Prolog syntax: Variables begin with capital letters, constants with lower case letters; the „,“ is the logical AND, and „?-“ initiates a query).

Example:

"Find the segmentation of the word "und" and display its labels in SAMPA".

```
?- word_canword(Id,und,_,_),
word_in_sentence(Id,_,SNo,WordPos),
segment_file(File,Spk,_,
  Sgr,SNo,_,Segs),
word_segments(WPos,Segs,WordSegs),
labels(WordSegs,WordLabels),
sampa_ipa(SampaLabels,WordLabels).
```

Complex applications combine the toolbox predicates with the standard control constructs of Prolog. The vowel duration analyses presented below are an example of a complex application; vowels are searched using a multi-level search pattern, e.g. "*vowel:, voiced-plosive". The code of the program proper is less than 20 lines, I/O and initialization require 15 lines each.

The PhDB contents are shown in table1

	number
word types	195
word tokens	676
segment files	5286
phonetic segments	238.769
reference segmentations	991
reference phonetic segments	39683

Table 1: PhonDat DB contents

The reference segmentations are the manual segmentations that have been selected for distribution within the PhonDat project. Phonetic segments do not contain para-phonetic (e.g. prosodic or syntactic) labels.

ANALYSES

Vowel Durations

Specific classes of speech sounds and the corresponding durations can be selected from the entire stored information by database queries only. For our investigations we chose the classic question about vowel duration and the influence of the following consonant and of vowel stress and length.

Stress and phonological length are factors that influence the duration of vowels [5]. In our analyses we compare the duration of stressed vs. unstressed vowels followed by a consonant and the duration of German long vs. short vowels in general and with voiced vs. voiceless following consonant.

Vowel duration is also very much influenced by the segmental context. In our study on 16 speakers we analyzed 9950 vowels that are followed by a consonant either within words or over word-boundaries. They are grouped according to the features stressed vs. unstressed; German central vowels /@/ and /6/ are looked at separately. Consonant classes are a) voiced/ voiceless, b) voiced/voiceless plosives, fricatives, and nasals. Our results of vowel duration measurements, which are shown in tables 2 - 6 provide further evidence to what is known from other investigations [6], [7]:

Phonologically long vowels have a longer duration than short vowels.

	number	duration
all V+C	9950	74
V long + C	3217	97
V short + C	6733	64

Table 2: number and average duration in ms of long vs. short vowels

Stressed vowels have longer duration than unstressed; the always unstressed "reduction" vowels /@/ and /6/ are shortest in duration.

	number	duration
stressed V + C	3710	96
unstressed V + C	6240	62
/@/, /6/ + C	1590	56

Table 3: number and average duration in ms of stressed vs. unstressed vowels and the German central vowel /@/ and /6/

Vowels before voiced consonants are longer than before voiceless consonants. This tendency is stronger with long vowels and not existent with short vowels.

	number	duration
V+C voiced	5595	76
V+C voiceless	4358	72
long V + C voiced	1334	115
long V + C voiceless	1883	84
short V + C voiced	4261	64
short V + C voiceless	1475	63

Table 4: number and average duration in ms vowels before voiced vs. voiceless consonants

When consonant classes are differentiated it appears that vowels have a longer duration before fricatives and nasals. This tendency is stronger when voiceless plosives and fricatives are considered separately.

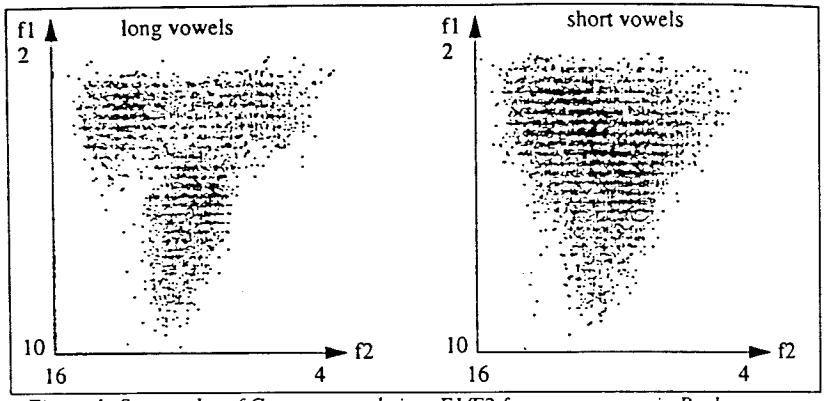


Figure 1: Scatterplot of German vowels in a F1/F2 frequency space in Bark

	number	duration
V + plosives	2586	84
V + fricatives	2557	67
V + nasals	3744	66

Table 5: number and average duration in ms of vowels before different classes of consonants

	V+C voiced		V+C voiceless	
	no	dur	no	dur
plosives	1056	89	1530	81
fricatives	88	90	2469	66

Table 6: number and average duration in ms of vowels before different classes of voiced vs. voiceless consonants

Formant Analysis

For the formant analysis we used the database tools to locate those sections in the speech signals that have been identified as vowels during the segmentation process. This information is used to apply a semi-automatic formant extraction program and measure F0 and the values of the first three formants of the vowels in PhDB. The program works as follows:

- The vowel segment plus 10 ms of the context is displayed in form of a spectrogram on a computer screen.
- A measurement point is proposed.
- F1, F2, F3 and F0 are extracted at the suggested time window in the signal.

The suggestion is based on the search for a local minimum of spectral variability as computed using cepstral difference coefficients [1]. The formant extraction is based on peak detection in a peak-enhanced 16 pole LPC-spectrum [8]. The F0-extraction is based on an autocorrelation PDA [4].

The formant values calculated by the program are marked and then checked by a phonetician who has the following options:

- accept the measurement
- select a new time location for the whole measurement by mouse click in the spectrogram
- correct the proposed formant values by mouse click at the preferred frequency in the spectrogram

The overall strategy to determine the measurement position was to find the target position of the vowel based on formant movement and energy. Altogether formant values of 10629 vowels have been extracted; reasonable F0 could be determined for 9228 vowels. Table 7 contains the average fundamental frequency and formant values of the analyzed vowels of PhDB.

Figure 1 shows the distribution of the spectral characteristics of long and short vowels in an F1/F2 scatterplot.

vowels	F0	F1	F2	F3
/i:/	185	324	2071	2698
/I/	175	369	1944	2698
/e:/	172	383	2076	2704
/E:/	176	443	2022	2660
/E/	169	486	1784	2633
/a:/	161	690	1339	2533
/a/	161	674	1362	2541
/O/	168	529	1162	2504
/o:/	165	416	927	2497
/U/	173	413	1093	2424
/u:/	199	328	946	2416
/y:/	174	341	1590	2298
/Y/	168	383	1541	2350
/2:/	177	395	1464	2249
/9/	168	477	1579	2291
/@/	178	449	1585	2570
/6/	173	535	1366	2490

Table 7: average F0 of 9228 vowels and formant values of 10629 German vowels (in Hz) for all 16 speakers of PhDB

Although the overall distribution is fairly similar, it appears by the degree of blackness that long vowels concentrate mainly in three regions: "back/round/high", "front/high" and "central/low". In contrast to this, the short vowels are distributed more regularly with a higher concentration in the centre of the vowel space.

CONCLUSION

Vowel duration and formant values of approx. 10000 German vowels have been measured. For duration measurements only the information stored in the Prolog database has been used. For formant measurement database information has been used to locate the exact position of vowels in the speech signal to apply a semi-automatic procedure for fundamental frequency and formant measurements.

Our findings of vowel duration and formant measurement support the results of earlier investigations [6], [7].

Our approach of combining symbolic database queries and acoustic analyses of speech signals has shown to be feasible and useful for phonetic research. The symbolic database has been extended with the formant data which is then available to further investigations.

The speech data of the PhonDat II train enquiry corpus is not phonetically balanced. The results we obtained may thus not be valid for spoken German in general. We plan to apply our methods to corpora with phonetically balanced data, and to larger speech corpora e.g. in cooperation with BAS [10].

REFERENCES

[1] van Bergem, D. R. (1993), "Acoustic vowel reduction as a function of sentence accent, word stress, and word class", *Speech Comm.*, vol. 12, pp. 1 - 23.

[2] Draxler, Chr. (1995), "Introduction to the Verbmobil-PhonDat Database of Spoken German", *Practical Applications of Prolog Conf. 95*, Paris.

[3] ECLiPSe (1992), *ECRC Common Logic Programming System*, User Manual, ECRC Munich.

[4] Hess, W. (1983): *Pitch Determination of Speech Signals*, Springer, Berlin.

[5] Lehiste, I. (1970), *Suprasegmentals*, Cambridge, Mass.: MIT Press.

[6] Maack, A. (1949), "Die spezifische Lautdauer deutscher Sonanten", *Zeitschr. f. Phon.*, vol. 3, pp. 190-232.

[7] Maack, A. (1953), "Die Beeinflussung der Sonantendauer durch die Nachbarkonsonanten, *Zeitschr. f. Phon.*, vol. 7, pp. 104 - 128.

[8] Markel, J.D., A., H. Gray (1976), *Linear Prediction of Speech*, Springer, New York.

[9] Pompino-Marschall, B. (1992), „PhonDat Daten und Formate", Institut für Phonetik, Uni München

[10] Tillmann, H. G. et al. (1995), "The Phonetic Goals of the New Bavarian Archive for Speech Signals", *Proc. of ICPhS*, Stockholm 1995.

CROSS-LANGUAGE VARIATION IN THE VOWELS OF FEMALE AND MALE SPEAKERS

Caroline Henton
Voice Processing Corp., 1 Main St., Cambridge, MA 02109, U.S.A.

ABSTRACT

After normalization, females' vowel spaces are uniformly larger than are males' spaces. Cross-language data indicate that female speakers produce more explicit vowels than do male speakers. It is particularly in the F1 dimension that the females' vowel quadrilaterals extend beyond the males'. It may be inferred that female speakers articulate vowels with a more open mouth. Acoustic and sociophonetic reasons for such behaviour are explored.

INTRODUCTION

Normalized acoustic data from seven languages and dialects indicate that female speakers produce vowels in a manner that is more phonetically explicit than that of male speakers. Against a background of acoustic ignoral of female-male differences, or sociolinguistic and dialectological inference, it is interesting to ask why it is that females are more 'open-mouthed' than are males, cf. [2],[4],[8]. Can we expect that speakers with a higher F0 automatically have a larger vowel space? Are females making a greater effort to keep vowels distinct, which might potentially contribute to greater intelligibility? Or do females over-articulate, avoiding reduced or centralized forms, as a result of social expectations to be 'guardians', and the overt wish to speak a prestigious variety of the language, see [8,9]?

Data from three English dialects will also be presented, showing that female speakers are not uniform in their behaviour: some females merge vowels more often than do males, while other female populations appear to differentiate the same vowels more systematically. The perceived social prestige of an accent is offered as one explanation for these disparate directions of change. Implications in speech technology, in terms of sex-differentiated recognition algorithms, and improved sex-specific speech synthesis (hypo- and hyper-articulation) will also be discussed.

CROSS-LANGUAGE DATA

Data from six phonetic studies are presented in Figure 1. Vowels are plotted in the F1-F2 space, with formants converted to the Bark scale. Data were normalized for perceptual comparison on a single referential system, by subtracting 1 Bark from the female values (1 Bark might equally well have been added to the male values). Motivations for this auditory normalization appear in [1], [6].

Details about the experimental collection methods for the languages and dialects appear in Henton, [4] and [6]. Languages illustrated in Figure 1 are (a) British English, Received Pronunciation (RP); (b) British English, Modified Northern (MN); (c) General American English; (d) French oral vowels; (e) Swedish long vowels; (f) Standard Dutch vowels. Vowels were also studied in Utrecht Dutch. From the series of plots, a pattern can be detected. In all cases, the females' vowel spaces are larger, more peripheral than those of the males, and particularly so in the F1 dimension.

DISCUSSION

Clear enunciation is a trait that has been associated consistently with female speech (see, *inter alia*, Kramer, [7]). This could mean several things phonetically. Firstly, women may 'over-articulate', i.e., they may use fewer grammatical and phonetic weak forms. From acoustic phonetic data it is impossible to observe grammatical hyper-articulation, since vowel measurements are most commonly obtained from word-lists or citation forms. It is nevertheless possible to speculate that women may produce phonetic hyperarticulation. Secondly, women may use the periphery of the articulatory space, compared with men whose vowels might be closer to the centre. In general, females' articulatory gestures appear to be more extreme across languages, and this greater articulation is achieved with the degree of jaw openness.

Both Labov [8] and Goldstein [2] have implied that, when possible, females will adopt a more 'open-mouthed' articulatory posture than will males. Labov [1990, p.219] returned to this theme more recently and observed that in some of the earlier literature research indicated that females were actively increasing the dispersion of the vowel system, by raising the peripheral tense vowels in Detroit English; whereas the male speakers exhibited shifts in the opposite direction, causing them to be more 'close-mouthed'. This tendency was also noticed tangentially for Swedish by Sundberg [10].

There is a possible connection between females' 'open-mouthedness' and the two principles of sexual differentiation in speech that Labov [8] invoked. Labov's first principle is that, "In stable sociolinguistic stratification, men use a higher frequency of non-standard forms than women" [1990, p. 205]. Linked to this principle is the fact that, "in (the) stable situations...women appear to be more conservative and favor variants with overt social prestige, whereas men do the reverse" [1990, p. 206]. There are a plethora of studies of various variables (notably the alternation of velar/alveolar nasal in "-ing/-in" in British, American and Australian English; the realizations of the interdental fricatives in English, and the various realizations of /s/ in Latin American and Peninsular Spanish) which all indicate that men use significantly more stigmatized forms than do women.

Any parallel between non-standard grammatical or lexical forms and the acoustic realizations of vowels has not, to my knowledge, been investigated explicitly, but would be worthwhile. Given that women can only exhibit their conservative or prestige-seeking behaviour when the opportunity arises, it does not seem unreasonable to assume that women in the phonetic studies were aware that they had been selected as speakers of a standard variety of the language; in the closely controlled environment of recording citation forms n a laboratory setting they would do their best to produce those standard and prestige forms that they had consciously or unconsciously come to guard.

The second principle proposed by Labov (ibid.) is that, "in change from below, women are most often the innovators." It is not possible to observe 'change from below' in the acoustic data here; in the one case where change might be said to have occurred - in the British English speech of the Modified Northerners (MN) - change has come from above, with the MN speakers changing their accents in the direction of the more prestigious RP.

Turning to the second of Labov's suppositions in his first principle above, namely that males' vowels are closer to the centre of a given vowel space, we may review data for two vowels in three dialects of English. The vowels are schwa and caret, which occur sequentially in an RP pronunciation of the word "above". Both are non-peripheral vowels. Using data from British English RP and MN, and West Coast American English, Henton [5] showed that in all three accents, the variability of the females' realizations of these vowels was significantly greater than that of the males. It was particularly in the F1 dimension that the females varied so widely, as could be seen in the coefficient of variation values. Furthermore, in West Coast American English, the male speakers centralized caret so much that, to all intents and purposes, they have only have one central non-rhotic vowel (schwa).

CONCLUSIONS

There is a regularity in the production of vowels across four languages, or seven dialects. Female speakers produced more open-mouthed variants of vowels than do males. If greater articulatory distinction may be equated with standard or prestige forms, then women can again be seen as guardians of the standard. Patricia Kuhl (personal communication) has indicated in her studies of cross-linguistic utterances by American English and by Swedish mothers to infants that the mothers tend to over-articulate, produce 'clearer' tokens when talking to babies than when talking to other adults (cf. Labov's reflections on the role of women in child-care, [1990, p.219].

To attribute linguistic change to one simple variable is dangerous. The exploration of these data invites an explicit investigation of whether women articulate more distinctively than men do. It has been suggested that re-plotting the current cross-language data on a

logarithmic scale would enable articulatory inferences to be made more appropriately than the Bark scale allows. A log. scale would probably render more conservative differences, but if (as is to be expected) the female-male differences remain, then the argument for females being more 'open-mouthed' would be all the more robust. Such a re-plotting will be presented in due course.

APPLICATIONS

Far from female speech being a "tongueless slobber" (Henry James, 1906), it seems that females make a greater to keep vowels more distinct than males do. With further data, it might be possible to determine empirically whether female speakers are more potentially more intelligible because of their greater differentiation of the peripheral vowels. For speech synthesis research this might imply that the F1 parameters of vowels in female speech need to be adjusted by a greater amount than might be expected by comparison to males' open vowel values. Intelligibility in speech synthesis has already reached asymptote, but increasing F1 would also add to the naturalness of the female voices. In speech recognition research it was assumed for many years that women were more difficult to recognize. The reasons given for such bias were nebulous, usually phrased in such terms as 'women's speech is so much more musical' or 'it's so variable'. Several facts militate against these sexist assumption: first, women's speech is not necessarily more intonationally variable (see [3]); second, current speech recognition techniques commonly dispense with any prosodic information; and third, women's vowels at least are more distinct than are mens'. Resistance has not lain in the technology, but rather been perpetuated by an androcentric scientific heritage focussing on males' speech alone. Most cogently, it has transpired that at least in English women's speech is generally *easier to recognize* than mens' speech.

Acknowledgements

A version of this paper was presented at the 123rd. meeting of the Acoustical Society of America, Salt Lake City, Utah. Thanks are due to Professors Marianna DiPaolo, Patricia Kuhl, and Terence Neary for their useful insights and suggestions.

REFERENCES

[1] Bladon, R.A.W., Henton, C.G. and Pickering, J.B. (1983) Towards an auditory theory of speaker normalization. *Language and Communication*, vol. 4, pp. 59-69.

[2]Goldstein, U. (1980) An Articulatory Model for the Vocal Tracts of Growing Children. Ph.D. dissertation, MIT, Cambridge, MA.

[3] Henton, C. (1995) Pitch dynamism in female and male speech. *Language and Communication*, 15, pp. 43-61.

[4] Henton, C. (1992) The abnormality of male speech. In G. Wolf (ed.) *New Departures in Linguistics*. New York, Garland Press, pp. 27-58.

[5] Henton, C. (1990) One vowel's life (and death?): the moribundity and prestige of /ɯ/. *Journal of Phonetics*, vol. 11, pp. 353-371.

[6] Henton, C. (1985) A Comparative Study of Phonetic Sex-specific Differences Across languages. D.Phil. thesis, University of Oxford.

[7] Kramer, C. (1978) Perceptions of female and male speech. *Language and Speech*, vol. 20, pp.151-161.

[8] Labov, W. (1990) The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, vol. 2, pp. 205-254.

[9] Labov, W. (1972) *Sociolinguistic Patterns*. Philadelphia, University of Pennsylvania Press.

[10] Sundberg, J. (1987) *The Science of the Singing Voice*. Dekalb, Ill., Northern Illinois University Press.

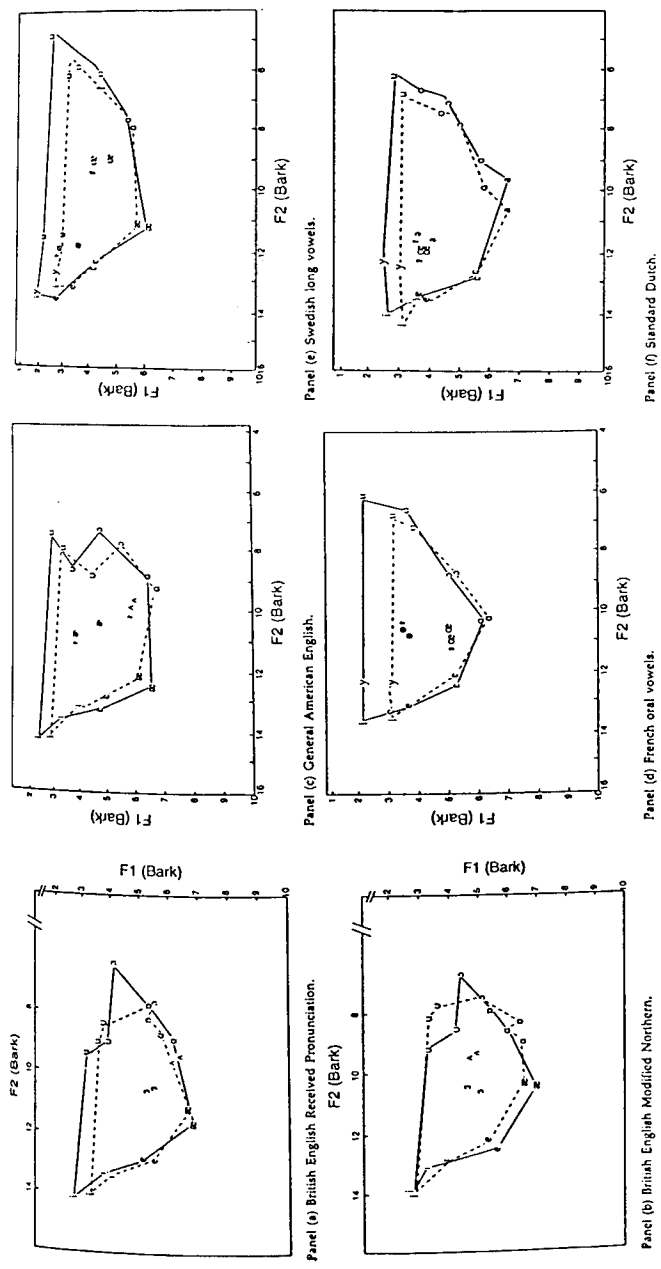


Figure 1 Mean formant values of vowels in each of six languages and dialects, plotted in a Bark-scaled F1-F2 space. Females' peripheral vowels are joined by a solid line, males' peripheral vowels by a dashed line. Panel (a) shows British English Received Pronunciation (RP); Panel (b) shows British English Modified Northern (MN); Panel (c) shows General American English; Panel (d) shows French oral vowels; Panel (e) shows Swedish long vowels; Panel (f) shows Standard Dutch vowels.

THE INFLUENCE OF THE PLACEMENT OF WORD-BOUNDARY ON THE ACOUSTIC INVARIANCE OF THE SYLLABLE

D.Horga
Dept. of Phonetics, University of Zagreb, Croatia

ABSTRACT

The influence of the higher language processing levels, namely the placement of the word-boundary (V1C#V2 or V1#CV2, e.g.tat#ući - ta#tući), on the syllabification patterns in Croatian is assessed. The consonants were /p/, /t/, /k/. Fifteen frame utterances were read by 10 subjects, recorded, and then digitised and treated to measure the following acoustic variables: the duration of V1, C, V2, and possible pause between C and V2, and the intensity of V1, V2 and consonant burst. The univariate analysis of variance and regression analyses show that C-closure duration , V2 and pause can predict the placement of the word-boundary.

INTRODUCTION

Discussing the phenomenon of production of smaller speech units some authors emphasize the importance of the phonetic level [1, 2, 3], the others insist on the conceptual and language level [4, 5, 6], while the third group suggest that the answer to the question should be "one and the other" and not "one or another" [7, 8]. Boucher [9] and Quené [10] found that the syllable acoustic parametars are influenced by the placement of the word boundary. Browman and Goldstein [11] proved the C-center can be the measurable parametar of the consonant belonging to the preceeding or following vowel. In the present paper the relationship between conceptual-language and phonetic levels of speech is examined by investigating the influence of the word boundary on the acoustic syllable structure in Croatian. Identical VCV segments which differ only in the placement of word boundary (V#CV or VC#V) are measured in frame utterances pronounced at a normal tempo and short enough not to require a syntactic pause within the VCV segment.

PROCEDURES

Fifteen pairs of sentences, 5 for each of the consonants /p, t, k/ were constructed. In the paired sentences the consonants were in the identical vowel context but the placement of the word boundary was either before or after the consonant. The sentences were matched according to the stress of observed vowels, number of syllables and rhytmic structure. For example:

Možda će ta tući. - a#tu
Možda će tat ući. - at#u

Ten female students of the Faculty of Philosophy in Zagreb, of normal speech and hearing status, read 30 randomized sentences. The sentences were recorded and then analysed by means of the computer speech program AGOS [12]. The investigated VCV syllables were described by measuring 8 acoustic variables: 5 variables of duration (first vowel - DV1, consonant closure - DCC, consonant burst - DCB, second vowel - DV2 and duration of the possible pause between consonant burst and second vowel - DPA), and 3 variables of maximal intensity (first vowel - IV1, consonant burst - ICB and second vowel - IV2).

The differences between V#CV and VC#V segments for each variable were tested by means of univariate analysis of variance. By means of multiple regression analysis the predictive strength of the variables to distinguish word boudary placement were determined.

RESULTS AND DISCUSSION

The results of the univariate analysis of variance are given in Table 1 and Figure 1.

The variables DCC, DV2 and DPA statistically significantly differentiate the two segments.

The consonant closure (DCC) is 16 ms (23%) longer if the consonant is placed after the word boundary (V#CV) than when it is before the boundary (VC#V). This result corresponds to that of Quene [10], who found that the consonant durations vary between 49 ms for intended CVC#VC, and 71 ms for intended CV#CVC, and post-boundary vowel rise time varies between 19 ms and 13 ms, respectively.

*Table 1. Means ($\bar{X}$) and standard deviations (s) of variables in ms in segments V#CV and VC#V. Statistically significant differences (p=.01) are marked by *.*

	V#CV		VC#V	
	$\bar{X}$	s	$\bar{X}$	s
DV1	68.6	28.9	68.5	28.1
DCC	85.3	19.4	69.5	16.5 *
DCB	23.8	15.6	26.0	16.9
DV2	115.1	26.0	124.3	24.8 *
DPA	0.5	6.0	36.9	33.6 *
IV1	62.5	2.6	62.3	3.1
ICB	49.8	5.3	49.3	4.8
IV2	62.5	2.8	62.5	2.7

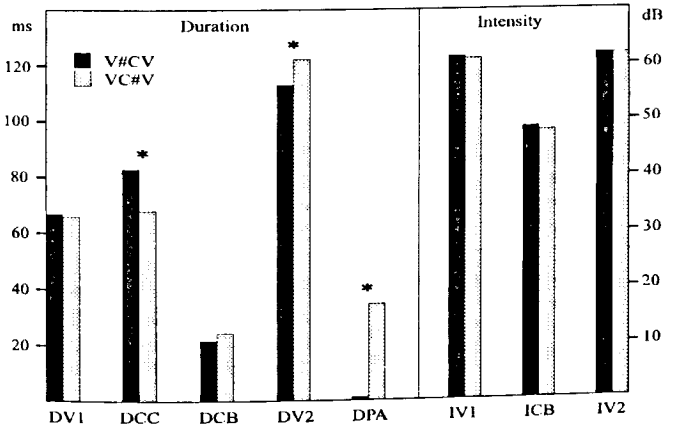


Figure 1. Means of measured variables in segments V#CV (black column) and VC#V (gray column).

It must be mentioned that closure in segment V#CV enables the speaker to produce a syntactic pause after the word boundary which cannot be separated from closure if the consonants are voiceless stops as in our experiment. Škarić [13] states that the average duration of syntactic delimitation pause equals approximately syllable duration, or about 100 ms. Our results show that the duration of consonant closure is shorter than that, so it could not include the syntactic delimitation pause. But it can be considered a kind of syllable delimitation or syllabic pause.

The duration of the second vowel in the segment VCV (DV2) is shortened in the position V1#CV2 by about 9 ms (9%) compared to the V1C#V2 position. This shortening can be explained by

faster articulation of the vowel when it is triggered by the consonant (V#CV) than when it is triggered by the intercostal muscles (VC#V).

Pauses between the consonant burst and the second vowel in the explored sample were found only in the V1C#V2 word boundary position. Of the 150 possible pauses in VC#V segments 96 (64%) were realised. Average duration of all the possible pauses was 40 ms; average duration of the realised pauses was 54 ms. As that is less than the duration of syntactic delimitation pauses, this kind of pause can be considered to indicate the syllabic structure.

Table 2. Multiple correlation (R), determination (R^2), regression coefficients (Beta) and correlations with the criterion (r).

	Beta	r	
DV1	.02	-.01	R = .67627
DCC	-.28 *	-.41	R^2 =.45734
DCB	.01	.07	
DV2	.15 *	.18	
DPA	.52 *	.60	
IV1	.01	-.04	
ICB	-.04	-.05	
IV2	-.12	-.00	

Regression analysis (Table 2) proved the results obtained by univariate analysis of variance. The chosen variables account for 46% of the variability of the whole system. The V1#CV2 and V1C#V2 were best positively predicted by the duration of the possible pause (DPA), duration of consonant closure (DCC) and negatively by the duration of second vowel (DV2). Such structure of the regression function shows that the segments V#CV and VC#V are best differentiated by the simultaneuos prolongation of the pause and V2, and shortening of the consonant closure, and vice versa. In other words, when in the VC#V segment the pause is realised at the word boundary the consonant closure is shortened. On the other hand, in the V#CV segment the zero pause is realised and the consonant closure is prolonged. These results prove the stability of the articulatory program and the possibilities of compensatory mechanisms at the level of sound articulation [8].

CONCLUSIONS

The investigation proved the influence of the placement of word boundary on some acoustic parametars of segments V1#CV2 and V1C#V2. The discrimination of the segments is mostly based on the duration of the syllabic delimitation pause and on the duration of the consonant closure, in which, potentially, the delimitation pause can be hidden. The duration of the second vowel was found to be less important element of discrimination. These variables are indicators of syllabic structure and the position of the word boundary. Intensity parameters did not prove to play a significant role in revealing the syllabic structure.

REFERENCES

[1] Koževnikov, V.A., Čistovič, L.A. (1965). *Reč'. Artikuljacija i vosprijatije*, Nauka. Moskva, Leningrad.

[2] Bondarko, L.V. (1984), *Fonetičeskoje opisanije jazyka i fonologičeskoje opisanije reči*, Izdatel'stvo Leningradskogo universiteta, Leningrad.

[3] Keller, E. (1991), "The distinction of the central and peripheral speech timing mechanisms", *Proceedings of the XIIth International Congress of Phonetic Sciences. Aix-en-Provence*, vol.1, pp. 6-9.

[4] Nooteboom, S.G. (1991), "Some observations on the temporal organization and rhythm of speech", *Proceedings of the XIIth International Congress of Phonetic Sciences. Aix-en-Provence*. vol.1, pp. 228-237.

[5] Carlson, R. (1991), "Duration models in use", *Proceedings of the XIIth International Congress of Phonetic Sciences. Aix-en-Provence*, vol. 1, pp. 243-246.

[6] Bell-Berti, F., Gelfer, C., Boyle, M., Chevrie-Muller, C. (1991), "Speech *Phonetic Sciences. Aix-en-Provence*, vol. 5, pp. 262-265.

[7] Kohler, K.J. (1991), "Isochrony, units of rhythmic organization and speech rate", *Proceedings of the XIIth International Congress of Phonetic Sciences. Aix-en-Provence*, vol. 1, pp. 257-261.

[8] Horga, D. (1992), "Varijabilitet govornih odsječaka", *Suvremena lingvistika*, vol. 34, pp. 81-92.

[9] Boucher, V.J. (1988), "A parameter of syllabification for VstopV and relative-timing invariance", *Journal of Phonetics*, vol.16, pp. 299-326.

[10] Quené, H. (1991), "Word segmentation in meaningful and nonsense speech", *Proceedings of the XIIth International Congress of Phonetic Sciences. Aix-en- Provence*, vol. 5, pp. 82-85.

timing in ataxic dysarthria", *Proceedings of the XIIth International Congress of*

[11] Browman, C.P., Goldstein, L. (1988), " Some Notes on Syllable Structure in Articulatory Phonology", *Phonetica*, vol. 45, pp. 140-155.

[12] Stamenković, M., Bakran, J., Miletić, M., Tancig, P. (1991), "AGOS - programski sistem za analizu govornog signala", In *Andrijašević, M., Vrhovac, Y. (Eds): Informatička tehnologija u primijenjenoj lingvistici*, DPLH. Zagreb, 1991.

[13] Škarić, I. (1991), "Fonetika hrvatskoga književnog jezika", In Babić, S., Brozović, D., Moguš, M., Pavešić, S., Škarić, I., Težak, S. *Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika*, HAZU, Globus, Nakladni zavod, Zagreb.

ACOUSTIC CORRELATES OF WORD STRESS AND THE TENSE/LAX OPPOSITION IN THE VOWEL SYSTEM OF GERMAN

Michael Jessen, Krzysztof Marasek, Katrin Schneider, Kathrin Clahßen
Institute of Natural Language Processing, University of Stuttgart, Germany

ABSTRACT

Acoustic correlates of word stress and the opposition between tense and lax vowels were measured in the speech of ten speakers of German. F1- or F2-frequency was found to be a significant tense/lax correlate across stress conditions and for most sets of vowels. Significant correlates of stress across tense/lax differences and vowel sets are vowel duration and closure duration.

1. INTRODUCTION

The difference in German between stressed and unstressed syllables and the difference between tense vowels as in *Schote* [o:] 'pod' and lax vowels as in *Schotte* [ɔ] 'Scot' is expressed in part by the same acoustic correlates. Vowel duration, for example, encodes both stress and tenseness [4]. Other correlates are largely specific to the expression of either stress or tenseness. For example, F0 expresses stress, but not tenseness [3]. This situation calls for an investigation in which stress and tenseness are varied independently in the stimulus material and measured for the same set of acoustic parameters.

2. METHOD

The following near-minimal pairs involving tense and lax vowels were selected, in which the segmental context was [t^h_l] throughout: *Ventil* [i:] 'valve' vs. *Tormentill* [ɪ] 'tormentilla', *Klientel* [e:] 'clients' vs. *Kartell* [ɛ] 'alliance', *Spital* [ɑ:] 'hospital' vs. *Metall* [a] 'metal', *Anatolien* [o:] 'Anatolia' vs. *Ayatollah* [ɔ] 'ayatollah', *Thulium* [u:] 'thulium' vs. *Schatulle* [ʊ] 'casket'. On the basis of each of the ten words both a variant with a stressed and one with an unstressed target vowel was triggered by appending the derivational suffixes *-isch* and *-ist*, respectively. For example, based upon *Klientel* and *Kartell* the following combinations of tenseness and stress were triggered (with stress marks added): tense stressed (*klientélisch*), lax stressed (*kartéllisch*), tense unstressed (*Klientélist*), and lax unstressed (*Kartéllist*). These four combinations are referred to as 'e-vowels'. Analogously, four i-, a-, o-, and u-vowels were triggered. The resulting 20 target words were read twice each by ten speakers of German, five female and five male. The recordings were digitized and analyzed acoustically. A number of different acoustic parameters were measured. The temporal parameters measured are closure duration of [t^h] (Clos), aspiration duration of [t^h] (Asp), vowel duration of the tense and lax target vowel (Vdur), and the duration of the following consonant [l] (Cdur). Onsets and offsets of F2, as well as the moment of stop release, served as the relevant events for the segmentation of these adjacent temporal intervals. Selecting F2-onset as the right-hand boundary of aspiration duration is motivated in [2]. The frequency of the first (F1) and second (F2) formant was measured half way into the target vowel. Measurements were also made of the mean F0 of the target vowel (F0mean), the standard deviation of F0 over the span of the target vowel (F0sd), as well as of the mean RMS (RMSmean) and standard deviations of RMS (RMSstd) over the vowel span. As another parameter, vowel energy (Energ) was calculated as Vdur * (RMSmean + 100) in analogy to a procedure proposed in [1].

3. RESULTS

For the presentation of the results and the statistical analysis the data of all ten subjects are pooled together, except for the F0-parameters, that are evaluated separately for female and male speakers. The measurement results for the durational parameters Clos, Asp, Vdur, and Cdur are represented graphically in Figure 1. No results for Clos in [u] are available because the [t^h] of the words *thulisch* and *Thulist* are not preceded by a segment (no more appropriate near-minimal pair could be found). The results

of the vowel formant measurements, as well as of the measurements of F0, RMS, and energy are presented in Table 1.

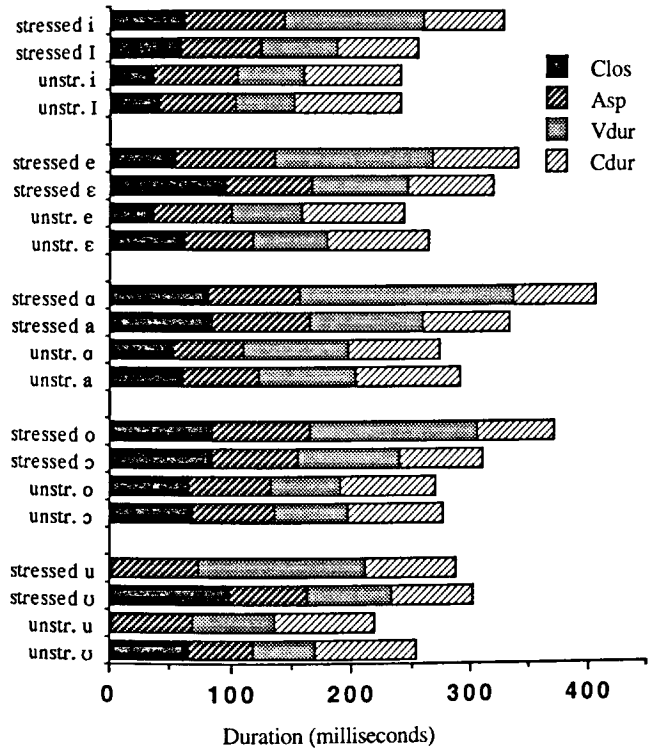


Figure 1. Mean results for the parameters Clos, Asp, Vdur, and Cdur pooled across the data from ten speakers. Duration values in milliseconds are presented horizontally, the conditions of tenseness and stress are presented vertically, divided into five separate blocks each for i-vowels (above) to u-vowels (below).

Table 1. Mean results for different acoustic parameters (horizontally) and different conditions (vertically), divided into five different vowel sets as in Figure 1. F1, F2, and F0 values are expressed in Hertz, RMS values are expressed in decibels.

target vowel	F1	F2	RMS mean	RMS sd	Energ	F0mean female	F0sd female	F0mean male	F0sd male
stressed i	299	2485	-20.8	1.1	9.0	218	5.6	149	5.0
stressed I	414	2150	-19.5	1.2	4.9	233	3.1	145	2.6
unstr. i	354	2369	-22.3	1.0	4.1	203	3.7	116	3.9
unstr. I	378	2192	-22.8	1.1	3.7	214	3.8	117	1.9
stressed e	381	2228	-19.6	0.9	10.5	231	11.8	129	5.6
stressed ɛ	569	1917	-17.8	1.5	6.5	217	8.6	129	2.3
unstr. e	407	2031	-22.2	1.1	4.4	207	8.8	115	4.8
unstr. ɛ	451	1878	-21.5	0.9	4.8	210	8.7	115	2.7

stressed ɑ	827	1375	-18.0	1.5	14.5	214	7.4	124	7.2
stressed a	807	1443	-18.0	1.7	7.6	215	7.6	118	2.4
unstr. ɑ	669	1606	-20.6	1.2	6.8	195	10.9	113	3.8
unstr. a	680	1590	-20.9	1.2	6.3	200	11.3	113	2.9
stressed o	363	738	-19.1	1.2	11.2	229	13.0	131	6.2
stressed ɔ	563	1086	-18.2	1.4	6.9	216	12.8	131	2.9
unstr. o	414	1138	-22.1	0.9	4.4	207	4.1	116	3.0
unstr. ɔ	453	1242	-21.9	0.9	4.7	203	9.4	118	2.9
stressed u	290	699	-19.7	1.2	10.9	226	5.8	143	6.9
stressed ʊ	406	1045	-18.2	1.2	5.7	224	3.7	145	2.7
unstr. u	354	1033	-21.3	1.3	5.2	199	3.9	122	2.5
unstr. ʊ	367	1207	-22.0	1.1	3.9	205	2.3	123	2.0

For the set of i-vowels four separate t-tests were calculated for each parameter. In the first t-test tenseness was the independent variable and the test was run on the stressed tokens, in the second it was run on the unstressed tokens, in the third t-test stress was the independent variable and the test was run on the tense tokens, and finally on the lax tokens. The same classes of t-tests was carried out on the other sets of vowels. Conditions for t-tests were proven correspondingly, and the data were found to meet the conditions. We present the statistical results by reporting only those parameters that were found to be significant (probability of t-statistics<.05). Table 2 lists for every set of vowels the parameters that were found to be significant in each of the four classes of t-tests.

Table 2. List of acoustic parameters that are significant according to t-tests in five different vowel sets (horizontally) and four different classes of t-tests (vertically) in the order mentioned in the text. Female and male are abbreviated as [f] and [m], respectively.

	i-vowels	e-vowels	a-vowels	o-vowels	u-vowels
tense vs. lax on stressed tokens	Asp, Vdur, F1, F2, Energ	Clos, Vdur, F1, F2, RMSstd, Energ, F0std [m]	Vdur, Energ, F0std [m]	Vdur, F1, F2, Energ, F0std [m]	Vdur, F1, F2, Energ, F0std [m]
tense vs. lax on unstressed tokens	no significant parameter	Clos, Asp, F1	no significant parameter	F2	Asp, Vdur, F2, Energ
stressed vs. unstressed on tense tokens	Clos, Asp, Vdur, Cdur, F1, Energ, F0mean [m]	Clos, Asp, Vdur, F2, Energ, F0mean [m]	Clos, Asp, Vdur, F1, F2, Energ, F0mean [m], F0std [m]	Clos, Vdur, Cdur, F1, F2, RMSmean, Energ, F0std [f], F0mean [m], F0std [m]	Vdur, F1, F2, Energ, F0mean [m], F0std [m]
stressed vs. unstressed on lax tokens	Clos, Vdur, Cdur, RMSmean, Energ, F0mean [m]	Clos, Asp, Vdur, Cdur, F1, RMSmean, RMSstd, Energ, F0mean [m]	Clos, Asp, Vdur, Cdur, F1, F2, RMSmean , RMSstd, Energ	Clos, Vdur, F1, F2, RMSmean, RMSstd, Energ, F0mean [m]	Clos, Asp, Vdur, Cdur, F1, F2, RMSmean, Energ, F0mean [m]

4. DISCUSSION

The results show that depending on the specific set of vowels involved and the tense/lax distinction, word stress in German is expressed by a variety of different correlates. The parameters that emerge from the results as the most reliable correlates of word stress, evaluated in terms of the occurrence of a significant effect across different conditions, are vowel duration (Vdur) and the duration of a stop closure (Clos) in the stressed vs. unstressed syllable. Since the energy index (Energ) includes Vdur, Energ patterns statistically with Vdur. The dependence of aspiration duration (Asp) on stress is not substantial and reliable enough to justify the existence of a phonological rule in German that assigns a feature category like [aspirated] to stops before stressed vowels (cf. [5]). Rather, the results suggest that aspiration depending on stress in German is a gradient phenomenon and part of phonetic implementation. The fact that consonant duration (Cdur) is smaller, while Clos, Asp, and Vdur are larger in the stressed than in the unstressed conditions indicates that lengthening due to stress is limited to the domain of the syllable (the consonant [l] belongs to the following syllable, which is unstressed if the preceding syllable is stressed). The F0 and intensity (RMS) parameters do not contribute with much reliability to the expression of stress. Although examination of the results for the individual speakers reveals that stressed vowels are consistently produced with higher F0mean than unstressed ones, the effect is not significant in several conditions. Significant effects for F1 and F2 in a number of conditions indicate that vowel quality is another correlate of stress in German. Among the set of tense vowels, vowels realized with stress are more peripheral in the vowel space than unstressed vowels. Among the lax vowels no clear similar nor opposite tendency can be observed.

In the evaluation of the different correlates of the tense/lax opposition in German it is important to realize that unstressed position imposes a strong constraint for the expression of the tense/lax difference. While, for example, vowel duration (Vdur) is significant across all stressed vowel conditions, it is significant for unstressed vowels only in the case of u-vowels. Formant structure (F1 or F2), on the other hand, is not only significant under stress, but remains significant in most conditions involving unstressed vowels. Similar results with an analogous set of target words are reported in [4]. As argued by [7], the stability of vowel quality across stress conditions speaks for the distinctive status of vowel quality, as opposed to vowel quantity in German. A-vowels behave differently from other sets of vowels. While all other vowel sets show significant effects for F1 and F2 in stressed position, a-vowels do not differ significantly in the expression of the tense/lax difference with respect to formant structure (cf. [6]). It is proposed in [4] that for low vowels (i.e. a-vowels) vowel quantity functions distinctively in German, while for nonlow vowels the distinctive property is vowel quality. A similar conclusion has been reached with results from vowel perception in German by [8]. We hope that further research will clarify why F0std for male speakers depends significantly on tenseness in most conditions involving stressed vowels.

REFERENCES

[1] Beckman, M.E. 1986. *Stress and non-stress accent.* Dordrecht: Foris.

[2] Davis, K. 1994. Stop voicing in Hindi. *Journal of Phonetics* 22: 177-193.

[3] Fischer-Jørgensen, E. 1990. Intrinsic F0 in tense and lax vowels with special reference to German. *Phonetica* 47: 99-140.

[4] Jessen, M. 1993. Stress-conditions on vowel quality and quantity in German. *Working Papers of the Cornell Phonetics Laboratory* 8: 1-27.

[5] Kloeke, W.U.S. van Lessen 1982. *Deutsche Phonologie und Morphologie.* Tübingen: Niemeyer.

[6] Ramers, K.H. 1988. *Vokalquantität und -qualität im Deutschen.* Tübingen: Niemeyer.

[7] Reis, M. 1974. *Lauttheorie und Lautgeschichte.* München: Fink.

[8] Weiss, R. 1977. The phonemic significance of the phonetic factors of vowel length and quality in German. *Phonologica 1976*, 271-276.

ACOUSTIC PROPERTIES OF NON-SIBILANT FRICATIVES

Allard Jongman and Joan A. Sereno
Cornell University, Ithaca, N.Y., U.S.A.

ABSTRACT

Two recent classification metrics, spectral moments and locus equations, were employed in an attempt to distinguish English labiodental fricatives /f, v/ from dental /θ, ð/. Preliminary results suggest that these two classes of fricatives are distinct, both in terms of spectral moments (primarily skewness and kurtosis) and slope and intercept of locus equations.

1. INTRODUCTION

A fundamental issue in speech research concerns whether distinctions in terms of place of articulation are more successfully captured by local (static) or global (and/or dynamic) properties of the speech signal. Most studies of place of articulation have investigated stop consonants (e.g., [1-3]). In contrast, fricatives are less well-known, and it is uncertain whether the classification metrics proposed for distinguishing place of articulation in stop consonants can be successfully applied to fricatives.

Although fricatives have been the subject of much research, the cues which serve to classify English fricatives according to place of articulation are still not fully understood. Acoustic studies focusing on the frication noise show that properties of the spectrum, amplitude, and duration of the noise can all serve to distinguish the sibilant /s, z, ʃ, ʒ/ fricatives from the non-sibilant /f, v, θ, ð/ fricatives (e.g., [4-6]). Within the class of sibilant fricatives, spectral properties serve to distinguish /s, z/ from /ʃ, ʒ/. However, none of the noise properties seems adequate to distinguish /f, v/ from /θ, ð/. Most research (e.g., [7]) suggests that acoustic cues to this distinction might be located in the fricative-vowel transitions.

The present study focuses on two recent classification metrics that, with appropriate modifications, seem particularly promising to investigate the role of these transitions as cues to the /f, v/ - /θ, ð/ distinction: spectral moments and locus equations.

Spectral moments analysis involves a statistical approach, capturing both local (spectral peak) and global (spectral shape) aspects of obstruents. Specifically, FFTs are calculated at different locations in the speech signal, and each FFT is then treated as a random probability distribution from which the first four moments (mean, variance, skewness and kurtosis) are computed. Mean and variance reflect the average energy concentration and range, respectively; skewness refers to spectral tilt and kurtosis is an indicator of the peakedness of the distribution. Previous research [8] using spectral moments has primarily examined the information derived from the first 20 ms of obstruent-vowel sequences. This approach reliably distinguished /s/ from /ʃ/, but failed to distinguish the non-sibilants from each other.

The locus equations approach is also statistical in nature. Locus equations are derived based on the second formant (F2) at vowel onset and at vowel midpoint (e.g., [9]). Locus equations constitute a dynamic representation of speech sounds since they express a relation between F2 at different points in the speech signal. Previous results indicate that the F2 starting frequency of a vowel preceded by an obstruent provides unique information about the articulatory configuration used to generate the consonant. Although locus equations have recently been successful in the classification of place of articulation in voiced stop consonants, researchers have only just begun to apply this method to fricatives. Recent research using locus equations to analyze fricatives has reported contradictory results. Fowler [10] shows consistently different slopes for /v/ and /ð/, while Sussman [11] does not.

In their present form, neither of the two approaches just described has been entirely successful at uniquely distinguishing (English) fricatives in terms of place of articulation. Specifically, none of the metrics is capable of reliably distinguishing /f, v/ from /θ, ð/.

Nevertheless, with appropriate modifications, spectral moments and locus equations seem particularly promising in capturing the defining properties of fricatives.

The present study extends the spectral moments approach by using a larger window size (40 ms instead of 20 ms) and by additionally examining possible cues present later in the frication noise and transition region. In addition, locus equations will be computed in an attempt to shed light on previous contradictory results. Since locus equations specifically encode information about F2 at vowel onset and vowel midpoint, they may provide a very appropriate metric to investigate the role of transition information.

2. METHODS

Three native speakers of American English (2 males, 1 female) were recorded in the Cornell Phonetics Laboratory. Targets were of the form CVC, with the first consonant being /f, v, θ, ð/, the vowel being /i, e, æ, ɑ, o, u/, and the last consonant always being /p/. Three repetitions of each of these targets were produced in the carrier phrase "Say ____ again".

All recordings were sampled at 22 kHz with 16 bit quantization using Waves+ software running on a SparcStation LX. Two types of spectral measurements were made. For spectral moments, FFT spectra were computed using a 40-ms full Hamming window at each of four locations: onset, middle, and offset of the fricative noise, and centered over vowel onset. For each window location, the first four spectral moments (mean frequency, variance, skewness, and kurtosis) were calculated. These moments were calculated from both linear and Bark transformed spectra.

For locus equations, LPC spectra were computed, using a 23.3 ms full Hamming window at two locations: vowel onset, and centered over vowel midpoint. Spectral peaks were picked from the LPC spectral displays.

3. RESULTS

a. Spectral moments

All four spectral moments were derived for each stimulus at four separate locations: onset, midpoint, and offset of the fricative, as well as centered over vowel onset. For each moment and window location, the moment data for the labiodental fricatives /f, v/ were contrasted to the dental fricatives /θ, ð/. Each analysis was performed on both the linear and Bark data.

At fricative midpoint, labiodentals can be distinguished from dentals both in spectral mean and skewness, with labiodentals showing a higher spectral mean and a more negative skewness. Additional differences in kurtosis are found at fricative onset and fricative offset, with labiodentals having more diffuse peaks compared to dentals. For the window location centered over vowel onset, there are significant differences in skewness and kurtosis, although at this location, labiodental fricatives show greater positive skewness and less diffuse peaks compared to dental fricatives.

Interestingly, the analysis of the Bark moment data show few differences distinguishing fricative place of articulation, except at the middle of the fricative noise, where all four moments show distinct differences between the fricatives /f, v/ and /θ, ð/.

b. Locus equations

Table 1 shows slopes and intercepts of the locus equations for each fricative for each of the three speakers.

Table 1. Summary of locus equation slopes and y intercepts (Hz) for each speaker for /f/, /v/, /θ/, and /ð/. F and M indicate female and male speakers, respectively.

	/f/		/v/		/θ/		/ð/	
Speaker	slope	y interc	slope	y interc	slope	y interc	slope	y interc
F1	.742	383	.645	542	.417	1083	.321	1301
M1	.759	373	.597	585	.592	660	.565	797
M2	.790	224	.779	273	.545	778	.575	789

As can be seen from Table 1, the two places of articulation have distinct slopes and intercepts. Labiodentals have high slope values and low intercepts while dentals have lower slopes and higher intercepts. Paired two-tailed t-tests confirm that the difference in slope is significant [t(5) = 4.80, p = .0049], as is the difference in intercept [t(5) = -5.66, p =.0024].

Figure 1 shows locus equation scatterplots for labiodentals (top) and dentals (bottom) for all three speakers. The regression line equation is $y=.728x+379$, $r^2=.95$ for the labiodentals and $y=.499x+894$, $r^2=.81$ for the dentals.

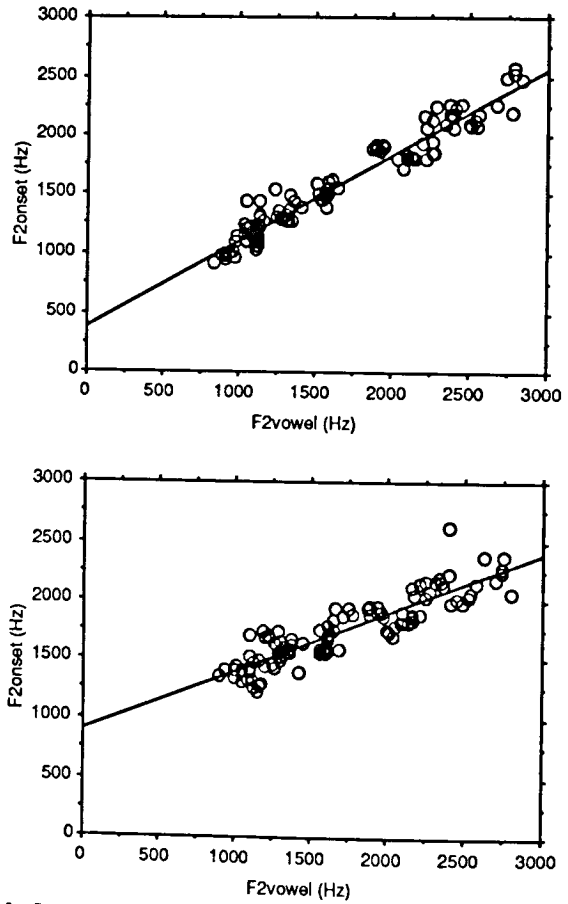


Figure 1. Locus equations for /f, v/ (top) and /θ, ð/ (bottom) for all six vowel contexts for all three speakers (two male, one female).

4. DISCUSSION

The present preliminary results show a number of interesting patterns. Spectral moments derived at the center of the frication noise and at the transition between noise and vowel seem most promising in distinguishing the non-sibilant fricatives in terms of place of articulation. The second moment, variance, does not seem to contain much distinctive information (cf. [8]). The use of the Bark scale does show some significant differences in place of articulation for the non-sibilant fricatives, but only when analyzing frication midpoint. Forrest et al. [8] found little effect of using non-linear transforms, perhaps because only onset information was examined. The present result suggests that location of the analysis window is crucial in determining place of articulation in non-sibilant fricatives.

Locus equations derived for /f, v/ and /θ, ð/ for three speakers are dissimilar. The two places of articulation seem to have distinct slopes and intercepts, with labiodentals having high slopes and low intercepts while dentals have low slopes and high intercepts. These present values are very similar to those reported for /v/ and /ð/ [10], both in terms of slope and intercept values. Although Fowler [10] questions the use of locus equations as cues to place of articulation across stops and fricatives, the use of locus equations as cues to place of articulation within each manner class is appealing since robust cues to the stop-fricative distinction are immediately available [12].

We feel the current results are encouraging. Both the moment analyses and the locus equations provide potential approaches for successfully distinguishing labiodental from dental fricatives in English. However, data for many more speakers are needed in order to conduct the necessary statistics (e.g., discriminant analyses for category discrimination both for the moment data and for slope and intercept values). A full dataset will be presented at the conference.

5. ACKNOWLEDGEMENT

We thank Scott Gargash for his technical expertise.

6. REFERENCES

[1] Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967), Perception of the speech code, *Psychological Review*, 74, pp. 431-461.

[2] Stevens, K.N., and Blumstein, S.E. (1981), The search for invariant acoustic correlates of phonetic features, In P.D. Eimas and J.L. Miller (Eds.), *Perspectives on the Study of Speech*, New Jersey: Erlbaum, pp. 1-39.

[3] Jongman, A. and Miller, J.D. (1991), Method for the location of burst-onset spectra in the auditory-perceptual space: A study of place of articulation in voiceless stop consonants, *Journal of the Acoustical Society of America*, 89, pp. 867-873.

[4] Hughes, G.W., and Halle, M. (1956), Spectral properties of fricative consonants, *Journal of the Acoustical Society of America*, 28, pp. 303-310.

[5] Strevens, P. (1960), Spectra of fricative noise in human speech, *Language and Speech*, 3, pp. 32-49.

[6] Behrens, S.J., and Blumstein, S.E. (1988), Acoustic characteristics of English voiceless fricatives: A descriptive analysis, *Journal of Phonetics*, 16, pp. 295-298.

[7] Harris, K.S. (1958), Cues for the discrimination of American English fricatives in spoken syllables, *Language and Speech*, 1, pp. 1-7.

[8] Forrest, K., Weismer, G., Milenkovic, P., and Dougall, R.N. (1988), Statistical analysis of word-initial voiceless obstruents: Preliminary data, *Journal of the Acoustical Society of America*, 84, pp. 115-124.

[9] Sussman, H.M., McCaffrey, H.A., and Matthews, S.A. (1991), An investigation of locus equations as a source of relational invariance for stop place categorization, *Journal of the Acoustical Society of America*, 90, pp. 1309-1325.

[10] Fowler, C.A. (1994), Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation, *Perception & Psychophysics*, 55, pp. 597-611.

[11] Sussman, H.M. (1994), The phonological reality of locus equations across manner class distinctions: Preliminary observations, *Phonetica*, 51, pp. 119-131.

[12] Jongman, A. (1989), Duration of fricative noise required for identification of English fricatives, *Journal of the Acoustical Society of America*, 85, pp. 1718-1725.

FORMANT TRANSITIONS: TEASING APART CONSONANT AND VOWEL CONTRIBUTIONS

S. Y. Manuel and K. N. Stevens
Research Laboratory of Electronics
Massachusetts Institute of Technology, Cambridge, MA USA

ABSTRACT

Acoustic data and models of consonant releases suggest that many CV transitions can be regarded as a sequence of two components: (1) an initial local change due to the release of the primary consonant articulator and (2) slower changes of F2 and F3 as the tongue body and jaw move away from positions as supporting articulators for the consonant constriction to their positions as primary articulators for the vowels.

INTRODUCTION

As is known, the transitions of the formants in consonant-vowel syllables reflect the identity of the articulator that makes the consonantal constriction, and where this articulator is placed. The transitions are also influenced by the articulatory configuration for the following vowel, and how much the vowel configuration is anticipated during the consonant constriction. (e.g., [1], [2], [3], [4], [5], [6]).

The release of a consonant in a symmetric VCV can be viewed as having two components, particularly when the primary articulator is the lips or tongue blade. The first component is the initial local movement of the primary articulator. The second, slower component consists of the movements of the tongue body and/or jaw, and possibly rounding of the lips, toward positions required for the vowel.

Here we explore the contributions of both components to formant transitions, and focus on labials and alveolars.

GENERAL METHOD

We are addressing this problem in two ways: (1) by calculation of formant movements for vocal tract models that are manipulated to change in a stepwise manner from a consonant-like to a vowel-like configuration; (2) by examination of natural speech. The use of modeling techniques augments the analysis of natural speech in several important ways. First, in natural speech, when the area of the constriction is still quite small, the sound may be dominated by a transient or frication burst created at the constriction, making it difficult to determine the natural frequencies of the vocal tract as a whole during this initial part of the release. Second, the initial formant movements can be quite rapid, and measurement is subject to the well-known problem of time-frequency trade-offs in accuracy. This problem is side-stepped with modeling which calculates formant frequencies for each step of the movement. Third, at present it is not possible to completely determine the vocal tract shape from the acoustic signal. Modeling allows one to be quite explicit about the vocal tract shape.

LABIAL STOPS

For labial consonants /b, p, m/, the lips alone can not form a constriction unless there is some jaw raising component, particularly with the jaw in a low position. Nevertheless, producing labials entails relatively little participation of the tongue body and jaw, compared with velar and alveolar consonants. Our preliminary modeling of labial releases suggests that, to a first approximation, the formant trajectories at the release of a labial stop in a symmetric VCV can be modeled by assuming a static vocal-tract shape with a time-varying cross-sectional area of the lip opening.

Using a computational model [7], we have emulated labial constrictions of various cross-sectional areas for several different tube shapes, and calculated the formants as the area at the lips is increased. The total tube length was 16.5 cm, divided into 33 sections that were each 0.5 cm long.

In Fig. 1 we show the calculated formant movements for labial release into /ɛ/. At release, F2 is about 1225 Hz and rises about 200 Hz as the area of the constriction increases to 0.5 cm^2. Assuming a rate of increase of opening

of about 50 cm^2/s, this increase in F2 would take place in the first 10 ms. As the opening increases, F2 continues to rise, though at a decreasing rate, and is still about 70 Hz short of its final value 40 ms after release. These modeled data conform reasonably well to formant values following the release of labials in the natural /ɛbɛ/ and /ɛmɛ/ utterances.

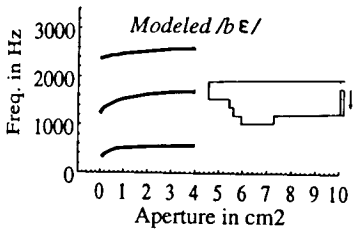


Figure 1. Labial before /ɛ/, showing model with varying aperture on the right, and calculated formants on the left.

For the vowel /i/, the movement of F2 is more extreme. For our modeled /i/ shape, at release F2 is highly sensitive to even very small changes in the labial constriction size, as shown in Fig. 2a. As the constriction area rises from zero to just 0.2 cm^2 (which should take only about 4 ms), F2 jumps from 1040 Hz to about 1800 Hz. F2 rises another 200 Hz as the area increases to about 0.5 cm^2, and then there is very little increase in F2 as the constriction opens further. In the modeled /bi/ the rise in F3 is quite a bit slower than the rise in F2.

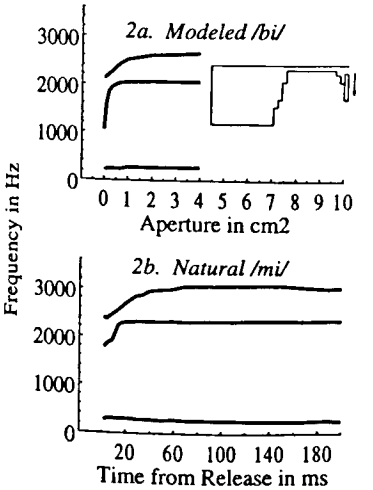


Figure 2. Labial before /i/.

In Fig. 2b we show formants measured at the release of /m/ in natural /imi/. Using nasals avoids the problem of interference of bursts, but still much of the initial rise is not observable as, explained earlier. The first measurable point for F2 was 2 ms after release, and had a value of about 1700 Hz. As in the modeled /bi/, F2 continues to rise another 200 Hz in the next 5 ms or so, and then the rate of rise slows down. F3 continues to rise after F2 completes its movement.

A different picture emerges for a shape similar to a back vowel such as /ɑ/ As indicated in Fig. 3, F2 is relatively flat following release. This is in contrast to the conventional wisdom (but see [4]) that F2 always rises coming out of a labial constriction. The lack of movement in F2 is presumably because the back cavity resonance does not change much during the labial release, and with even a moderate labial constriction, F2 is usually a back cavity resonance. Formant measures made at the release of natural tokens of /ɑbɑ/ and /ɑmɑ/ show a slight rise of F2, and modeling suggests that this is due to tongue body or jaw movement.

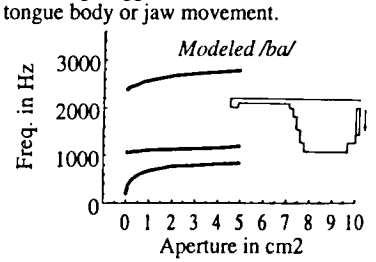


Figure 3. Labial before /ɑ/.

This work is consistent with earlier results (e.g. [4]) for labial releases, and suggests that the general pattern of changes in F2 right after release can probably be attributed to changes in lip aperture, which are superimposed on, and dominate, smaller tongue body and jaw effects.

ALVEOLAR STOPS

We turn now to alveolar stop constrictions (such as /d, t, n/) made with the tongue blade. Due to

anatomical constraints, the tongue body must be fronted to allow the tongue blade or tip to make an alveolar constriction. If an alveolar consonant is followed by a back vowel such as /ɑ/, the backing movement of the tongue body should produce a falling F2, and for a following /i/, F2 should rise as the body moves up and forward. However, the initial release of a constriction in the alveolar region should, in theory, result in an initial *rise* of F2. Therefore, one might expect to see a two-part formant trajectory at the release of an alveolar into a back vowel. First one should see an increase in F2, and then a decrease as the tongue moves back.

To examine this expectation, we began our modeling with the vowel /ɛ/, as we expect that in making the alveolar constriction in an utterance like /ɛdɛ/ there is very little tongue body adjustment, with the primary change in vocal tract shape being due to raising the tongue tip. This provides us with an idea of what the tongue-tip constriction itself contributes to the overall formant trajectory. In Fig. 4a the solid lines show what happens to a basic /ɛ/ shape when a constriction of various sizes is made 2 cm back from the front of the tube. The length of the aperture is 0.5 cm. When the aperture is zero, F2 is as low as 1380 Hz, then rises to about 1540 Hz by the time the aperture is 0.3 cm^2. This amount of change in F2 seems large, given the patterns seen in natural speech. We remodeled this constriction with a more anatomically correct tapered tongue tip, and the results are indicated in Fig. 4a as lines with circles. In this case, F2 is only as low as 1590 Hz at release. The raising of F2 when tapering accompanies the tongue-tip constriction is expected from perturbation theory, which predicts an increase in a formant frequency when a narrowing is made near a pressure maximum. The overall pattern is one of a slight rise in F2. We note that all of this movement can be attributed to tongue tip changes, as the tongue body shape was fixed. The natural /nɛ/ in Fig. 4b also shows little F2 movement.

We model /di/ much as we do /dɛ/, with the exception that the emulated tongue-body constriction is tighter and more forward for the /di/. Changing

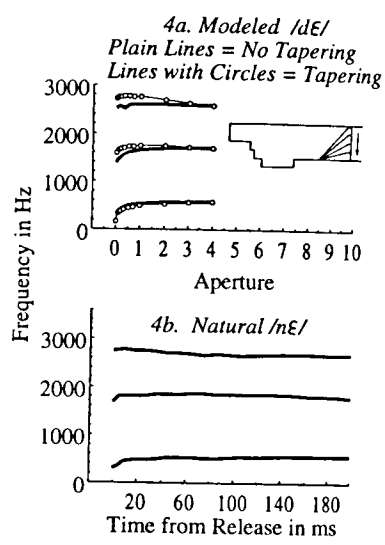


Figure 4 Alveolar before /ɛ/.

tongue-tip aperture alone results in a very rapid increase in both F2 and F3 as the aperture increases from zero to 0.5 cm^2, as shown in Fig. 5a. Not shown is the fact that tapering the tongue tip increases F2 and F3 substantially when the constriction is quite small. This may account for the fact that in natural speech, F2 and F3 do not appear to be particularly low at onset of the formants for /di/ and /ni/. In any case, whether due to actual differences in formant values, or to measurement problems, F2 may be quite similar for /di/ and /bi/. However, F3 is rather different for the two syllables, rising rapidly after the alveolar, but slowly after the labial.

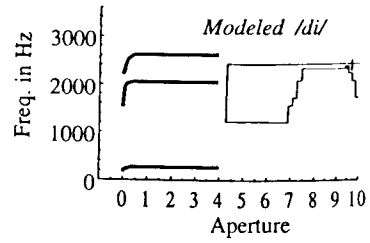


Figure 5. Alveolar before /i/.

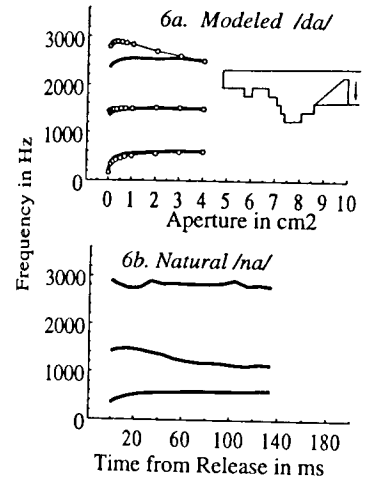


Figure 6. Alveolar before /ɑ/.

For /dɑ/ we begin with a shape similar to that of /ɛ/, but slightly more open in front and more constricted in the back. From this basic shape we adjust the tongue tip constriction, including the tapering, as noted above. As shown in Fig. 6a, there is very little movement in F2 - only a small rise. Of course, one normally expects to see a decrease in F2 as one moves from /d/ to /ɑ/. This model shows that such a drop in F2 must be due to the tongue body moving back, and not to the change in the consonant constriction itself. In the natural utterance /nɑ/ shown in Fig. 6b one can see a two-part F2 transition: a rather short, flat part followed by a longer, falling part. Modeling suggests that the former portion is due to a combination of the release of the alveolar constriction and some tongue backing, and that the latter portion is simply due to tongue backing.

SUMMARY

Formant transitions in the first 20-odd ms following labial releases are primarily a consequence of lip movement, and are probably not greatly influenced greatly by movements of the tongue body. Even after this initial interval, the influence of tongue-body movement is not very great, as the tongue body does not move much.

There are substantial differences in F2 movements following labial releases into front vowels like /ɛ/ compared with back vowels like /ɑ/. There is very little F2 movement for back vowels because the back cavity resonance does not change appreciably during the release. For front vowels there is a significant and rapid upward movement of F2, and a more slowly rising F3.

Formation of an alveolar constriction requires tapering of the vocal-tract area for a few cm posterior to the point of contact. With this tapering, the upward movement of F2 following release of an alveolar constriction is minimized.

Following an alveolar release into a front vowel like /ɛ/, the tongue-body movement is small, and as a consequence of tongue-blade tapering, there is little F2 movement. Following an alveolar release into a back vowel, the combination of the tapered constriction and the backward tongue-body movement leads to an initial flat F2 trajectory followed by a slow downward movement.

ACKNOWLEDGMENT

Work supported in part by NIH grant DC00075.

REFERENCES

[1] Delattre, P. C., Liberman, A. M., & Cooper, F.S. (1955) Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Am., 27,* 769-773.

[2] Fant, G. *Acoustic theory of speech production.* The Hague: Mouton.

[3] Stevens, K. N., & House, A S. (1956) Studies of formant transitions using a vocal tract analog. *J. Acoust. Soc. Am., 28,* 578-585.

[4] Fujimura, O. (1961). Bilabial stop and nasal consonants: a motion picture study and its acoustical implications. *J. Speech & Hearing Res., 4,* 233-247.

[5] Ohman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *J. Acoust. Soc. Am., 39,* 151-168.

[6] Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. *J. Acoust. Soc. Am., 90,* 1309-1325.

[7] A program developed by S. Maeda was used to make the computations.

3-D FEM ANALYSIS OF VOCAL TRACT MODELS USING ELLIPTIC TUBES WITH VOLUME RADIATION

Hiroki Matsuzaki, Nobuhiro Miki and Yoshihiko Ogawa
Faculty of Engineering, Hokkaido University

ABSTRACT

Using a 3-D FEM, we compute the sound pressure and the particle velocity in our vocal tract model with a volume radiation in order to estimate the 3-D effect. From the simulation results, we were able to show the differences between 1-D and 3-D models for the formant frequency and band width in the following case: the cross section of the tube is flattened from circle to ellipse, and the shape of constriction is complex.

1 INTRODUCTION

The shape of the vocal tract and the aperture at the lips are important factors in characterizing speech sound . A vocal tract model with cascading elliptic tubes was used to represent the 1-D equivalent circuit model. In this model, however, a plane wave is assumed on the radiational part, and not only the 3-D effect of radiation, but also the 3-D effect of the constriction of the incisors is neglected. We showed that there was a 3-D effect in the elliptic tube by using a 3-D FEM[1]. In this paper, using the FEM, we try to evaluate the 3-D effect of the radiation and the constriction of the incisors. We compute the sound pressure and the particle velocity in our vocal tract model including the volume radiation. The volume radiation is hemisphere shaped. From the experimental results, we see that the 3-D effects are large on the Vocal Tract Transfer Function(VTTF). Finally, we propose a new vocal tract model with cascading non-uniform elliptic tubes. This model is based on MRI data of vocal tracts, and the shape of several cross sections is determined by the elongation factor and the area.

2 FORMULATION OF THE WAVE EQUATION

It is well known that the acoustic wave equation in steady state is represented using velocity potential ϕ as

$$\nabla^2\phi = k^2\phi \quad (1)$$

where $k(\omega$(angular frequency)$/c$(sound velocity)) is the wave length constant, and that sound pressure p and particle velocity $\boldsymbol{v}$ are represented as

$$p = j\omega\rho\phi \quad (2)$$

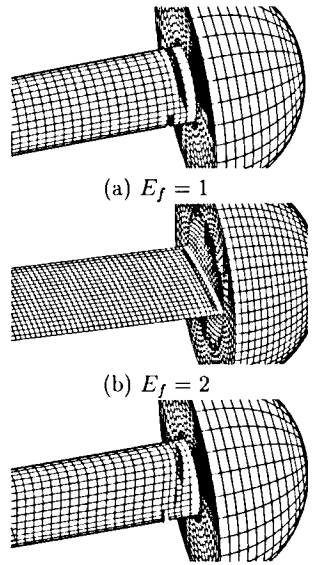
$$\boldsymbol{v} = -\nabla\phi \quad (3)$$

where ρ is the atmospheric pressure density. Our FEM formulation was based on the above equations.

3 SIMULATIONS FOR CONSTRICTION

In order to evaluate the effect of the constriction of the incisors in vocal tracts, we computed acoustic characteristics in some straight sound tubes with the constriction. We use a simulation model of the sound tube with a cross sectional area of πcm^2, and a length of $15cm$. The two cross sectional shapes of the tube were determined by a parameter E_f(Elongation Factor)[2]; E_f=1 for a circle and E_f=2 for an ellipse. Its driving surface is driven by sound velocity $v_n = 1e^{j\omega t}$. A volume of radiation, which is hemisphereical in shape, is attached to the aperture surface[3]. The radius is $3cm$ for E_f=1 and $4cm$ for E_f=2[4]. As a boundary condition on the volume, the specific acoustic impedance of spherical radiation is assumed on the spherical surface, and a rigid wall baffle is assumed. The walls making constrictions

are shaped by half of the cross section with a thickness of $0.17cm$ and are located in the upper and lower parts of the tube. The lower one is at a distance of $14.25cm$ from the driving surface and the upper one is at a distance of $14.85cm$. In Fig.1, we show the close-up figures of our finite element models with a circular cross section(E_f=1) in (a), and an elliptic cross section(E_f=2) in (b). Moreover, in the case of E_f=1, the constrictions are rounded as to approximate the shape of the incisors(See Fig.1(c)).



(a) $E_f = 1$

(b) $E_f = 2$

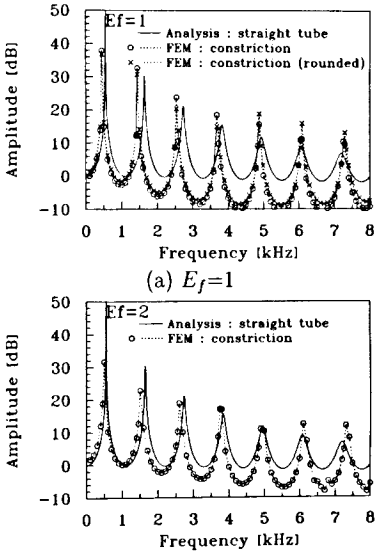
(c) $E_f = 1$: Rounded

Figure 1: *Finite element model*

In Fig.2, we show the VTTF computed from our FEM and the 1-D analytical model without incisors . From the particle velocity, simulated by the FEM analysis, the VTTF $H_v(\omega)$ is computed as

$$H_v(\omega) = 20\log_{10}\left|\frac{\sum v_l(\omega)/A_l}{\sum v_g(\omega)/A_g}\right| \quad (4)$$

where v_g (v_l) are the normal component of particle velocity at the driving (the aperture) surface, and A_g (A_l) is the area of the driving (the aperture) surface. In the results of our FEM



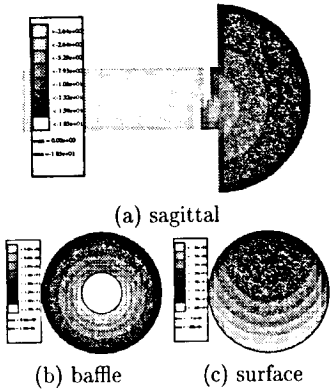
(a) E_f=1

(b) E_f=2

Figure 2: *Vocal tract transfer function*

model, up to the fourth (or fifth) formant, the formant frequencies shift to lower frequencies. This means that the acoustic length of the tube seems longer than the real length. The differences of the formant frequencies between the non-rounded and rounded constriction models are a few Hz. In the case of E_f=2, there are some peaks on the VTTF between $6k$ and $7kHz$;and these peaks are discussed in the following experiment.

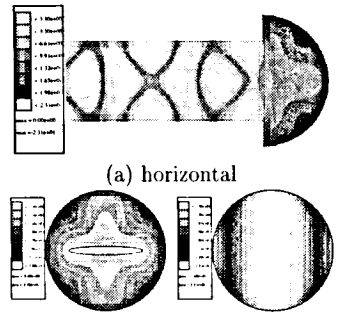
Fig.3 and 4 show the sound pressure distributions([dB]) for E_f=1(frequency is $1450Hz$) and E_f=2(frequency is $6800Hz$). Fig.3(a) is on the sagittal plane and Fig.4(a) is on the horizontal plane. (b) of both models is on the baffle of the volume, and (c) on the spherical surface. In Fig.3(a), we see that the sound wave propagates along



(a) sagittal

(b) baffle (c) surface

Figure 3: *Sound pressure distribution* ($Ef = 1,1450Hz$)



(a) horizontal

(b) baffle (c) surface

Figure 4: *Sound pressure distribution* ($Ef = 2,6800Hz$)

the acoustic pass of the constriction. If the analytical model is used with the equivalent length and area of the tube estimated from the FEM results, the VTTF may be in agreement with the FEM results. In Fig.3(b) and (c), we see that the sound wave propagates non-symmetrically with respect to the upper and lower direction because of the alternate constriction. In Fig.4(a), we see that a higher mode is formed. In Fig.4(b), the sound wave does not always propagate along the aperture. In Fig.4(c), the sound wave is distributed vertically, therefore we guess that the sound wave propagates in a vertically polarized wave through free space.

4 SIMULATION OF A NEW VOCAL TRACT MODEL

The traditional vocal tract model is modeled by cascading uniform circular tubes. In this model, the tubes are not connected smoothly, although the outline of the human vocal tract is smooth. The question then arises about the non-smooth connection. In this simulation model, we connect the elliptic tubes smoothly. Our model is based on MRI data of the vocal tract for the Japanese vowel /a/[5], and the shape of several cross sections is determined by E_f and the area. In Fig.5, we show a finite element model including a volume radiation (radius of $4cm$).

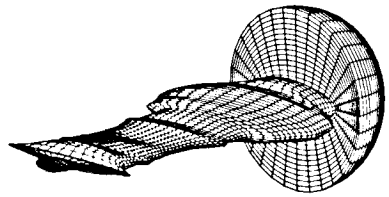


Figure 5: *Finite element model*

In Fig.6, we show the VTTF computed from our FEM and the traditional 1-D analytical model. The first and second formant frequencies of the FEM solution agree relatively with analytical solutions, but the third formant

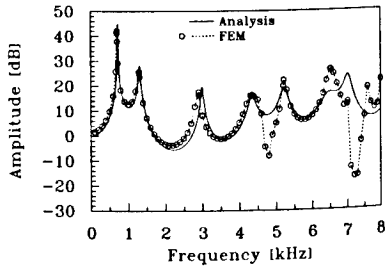
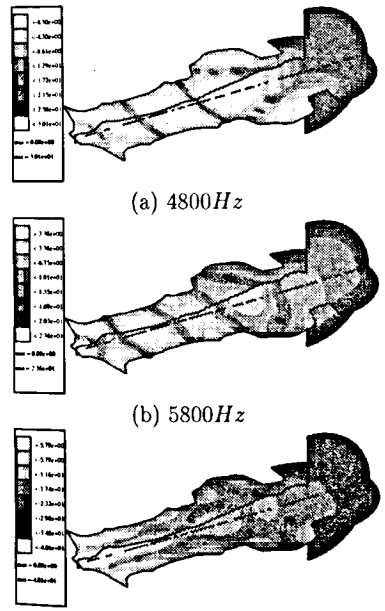


Figure 6: *Vocal tract transfer function*

of FEM is about $100Hz$ lower than the analytical one. In the higher formant frequencies, there are difference in the formant frequency or band width. And there are two valleys(zeros) between $4.5k$ and $5kHz$ and between $6.8k$ and $7.6kHz$.

In Fig.7, we show the sound pressure distributions($[dB]$) on the sagittal and the horizontal plane. The driving frequencies of $4.8k$ and $7.2kHz$ correspond to the two valleys(zeros), and $5.8kHz$ to the non-zero. In the large



(a) $4800Hz$

(b) $5800Hz$

(c) $7200Hz$

Figure 7: *sound pressure distribution*

cavity corresponding to the oral cavity, the effects of the higher mode in Fig.7(a) and (c) are larger than in (b).

CONCLUSION

Using our 3-D FEM models, we simulated the sound wave propagation in a vocal tract model with the constriction of incisors and in our model for the Japanese vowel /a/. These models included the volume radiation. The VTTF was computed by simulated result. And the sound pressure distributions were shown.

The results of the experiment showed 3-D effects on the formant frequencies because of the 3-D shape of the vocal tract with constriction and the existence of higher modes. These facts never appear in the traditional 1-D analysis. We consider that these facts are useful for natural speech analysis.

ACKNOWLEDGEMENT

We are grateful to Dr. Naohisa Kamiyama for his offer of valuable data and Dr. Kunitoshi Motoki for valuable advice.

REFERENCES

[1] Matsuzaki,H., et al. (1994), "Analysis of acoustic characteristics in the vocal tract with inhomogeneous wall impedance using three dimensional FEM model", Electron.& Commun. Jpn.

[2] Kamiyama,N., et al. (1991), "A study on the effect of the viscoelastic vocal tract wall", Proc. Fall Meet. Acoust. Soc. Jpn., 1-6-12, pp.213-214(in Japanese).

[3] Matsuzaki,H., et al. (1994), "3D FEM analysis of vocal tract model of elliptic tube with inhomogeneous-wall impedance", ICSLP94, S12-17,Vol.2, pp635-638.

[4] Motoki,K. & Miki,N.(1994) "Distribution characteristics of acoustic impedance density around the radiation area", JASA, 1-8-8, pp.643-644(in Japanese).

[5] Kamiyama,N., et al. (1992),"Study of the vocal tract impedance using viscoelastic model of the wall", Jpn. IEICE Trans(A), **J75-A**, 11, pp,1649-1656(in Japanese).

PITCH DEPENDENCY OF VOCAL TRACT TRANSFER FUNCTIONS

Nobuhiro Miki[*],
Pierre Badin[**], *Kenji Takemura*[*], *Masahiro Kuroda* [*] *and Yoshihiko Ogawa*[*]
[*]*Faculty of Engineering, Hokkaido University*
[**] *Institute de la Communication Parlee*

ABSTRACT

The variation of the vocal tract transfer function (VTTF) is discussed for the vowels with pitch variation. Even in the isolated vowels, we show the fluctuation of frequency and bandwidth in the higher formants of the upper third formant; this fluctuation pattern corresponds to small pitch variations. We try to compare the VTTF variation between Japanese and Swedish speakers.

1 INTRODUCTION

It is well known that the accent in Japanese speech is controlled by pitch variation. In the traditional theory of speech production, it has been assumed that the vocal tract transfer function (VTTF) is not changed by pitch variation in steady state articulation. Recently, however, it is known that the position of the vocal fold moves up or down a few cm in pitch control for Japanese vowels even in steady state articulation [3]. If the size of the vocal tract is varied in the order of cm, the corresponding transfer function should be varied in the phonation of steady state vowels. From this point of view, we tried to estimate the VTTF variation from the speech signal by using our short-time speech analysis algorithm (M-algorithm)[1], and we pointed out that this variation is associated with the variation of the 3-dimensional figure of the laryngeal part caused by pitch control [2]. In order to estimate the VTTF from the cut speech signal corresponding accurately to the glottal closure, we record speech signals and EGG signals simultaneously using a DAT. From the estimated results, we show the fluctuation of formant frequencies even in the isolated vowels or the accentuated vowels. Moreover we try to compare the VTTF variation between Japanese and Swedish speakers.

2 DIRECT MEASUREMENT OF VOCAL TRACT TRANSFER FUNCTION

Our interest is to find the relation between the laryngeal movement including the glottal height and the variation of VTTF, and we used the VTTF measurement system of ICP [4]. In this system the shaker on the neck is driven by a white noise sequence; the response at the lips is picked up by a microphone, and is digitized by a PC. We compare the difference in the estimated VTTFs of the vowel under the two conditions of glottal control;

(a): The subject utters a high pitch vowel, closes the glottis for few seconds, utters again with low pitch, and then closes the glottis. The subject keeps the articulation as neary the same as possible during this measurement. The VTTF is measured in two parts of the glottal closure, and the difference is compared for high and low pitch.

(b): The subject utters a vowel, closes the glottis for few seconds, and then relaxes the glottis. The VTTF is measured in the closure part and the relaxed part.

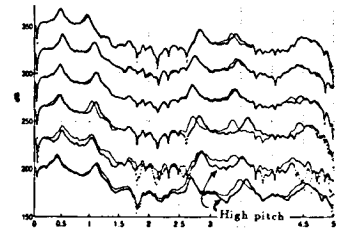


Fig.1 Results for experiment (a)

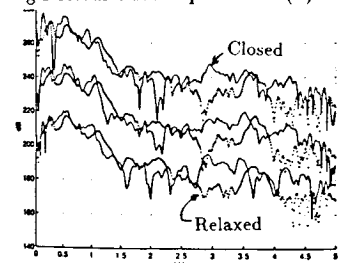


Fig.2 Results for experiment (b)

The experimental results are shown in Fig.1 and 2. As in Fig.1, the difference of VTTFs appears, but is smaller than that of Fig. 2. In the relaxed condition the position of the glottis is lower than in ordinal phonation, and we see the influence of the subglottal system on the VTTF. In the above experiments, the measured VTTFs are average frequency responses in the several hundreds mm seconds under steady state glottal conditions. We see that the higher formants shift to high frequency directions according to height position condition of the glottis in both figures.

From these data, we see that the VTTF (formants) can be varied even in the same vowel by pitch control.

3 EVALUATION OF THE VARIATION OF VTTF BY USING A MODEL

In order to evaluate the formant shift with the movement of glottal position, we compute the two kinds of VTTF for high or low pitch by using the vocal tract model and the 3-D data of the vocal tract from MRI data. The difference of VTTF is shown in Fig. 3; the glottis moves up 0.5 cm and the laryngeal part near the glottis shrinks about 15 % . From this figure we can conclude that the formant shift is larger in the higher formants, but the shift frequency is small.

We can suppose that during utterance the glottal movement is from a relaxed position to an upper position, and back again to the relaxed position. Since this small movement influences the VTTF and gives a dynamic shift on the formants, we need to evaluate this dynamic shift by using a short-time analysis algorithm.

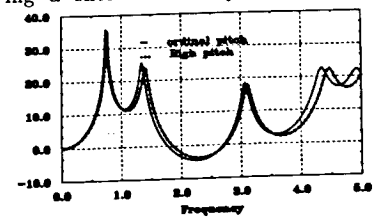


Fig.3 VTTFs computed by the model

4 ESTIMATION OF THE VARIATION OF VTTF IN ISOLATED VOWELS

Using a detection algorithm of peaks of the EGG signal, we obtain a

short-time interval data cut from the closed phase of vowels. Since the obtained data have discontinuous points, the data are mapped to continuous waveforms with the Fejer kernel, and are analyzed to find accurate VTTFs (formants) by using the M-algorithm [1]. As in the first experiment, the subject uttered a vowel with pitch control from low to high frequency during 400 - 600 mm sec. In each glottal closure, the VTTF was estimated with the M-algorithm. Fig.4 shows the difference of VTTF of /a/ uttered by a Japanese speaker, and we see that almost all the formants shift from low to high frequency according to the pitch frequency. In Fig.5, the difference is shown for a Swedish speaker; the frequency shift of formants is smaller than in Japanese. It may be caused from different articulation for /a/.

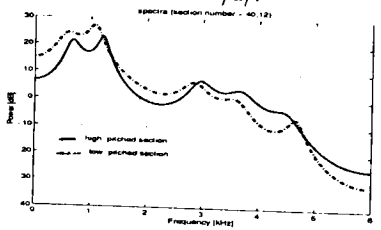


Fig.4 VTTF of Japanese /a/

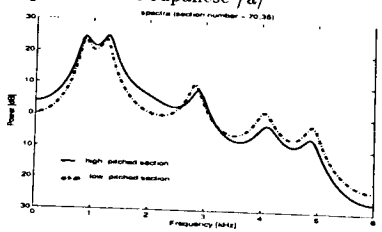


Fig.5 VTTF of Swedish /a/

As the second experiment, we estimate the formant movement for the Japanese vowel /a/ with pitch control from high to low frequency, and the result is shown in Fig.6 (a) and (b). The estimated formants are not of steady state frequency but of variable pattern. Around 2 kHz we see the false formant sequence which may be caused by the subglottal coupling (incomplete closure), and in this duration the 5th formant of the upper 4kHz range is blurred. The estimated formants vary with pitch frequency, and the subglottal coupling appears in some durations. From these experiments, we see that the VTTF can be varied with pitch frequency even in the isolated vowels.

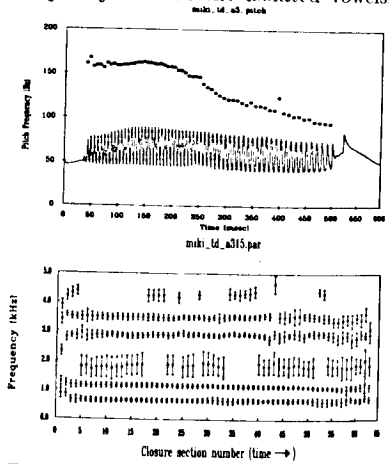


Fig.6 Pitch frequency and traced formants

5 ESTIMATION OF THE VARIATION OF VTTF IN VCV

It is interesting to compare the two VTTF for vowels in the sequence VCV (Vowel Consonant Vowel). Here we show the estimated VTTFs for the two /a/s in /aká/ uttered by Japanese and Swedish. Although the these VTTFs are influenced by the articulation of /k/,we see the formant shift from the first vowel to the second vowel; the formant shift for Japanese is larger than for Swedish, and the influence of /k/ for Japanese is also larger.

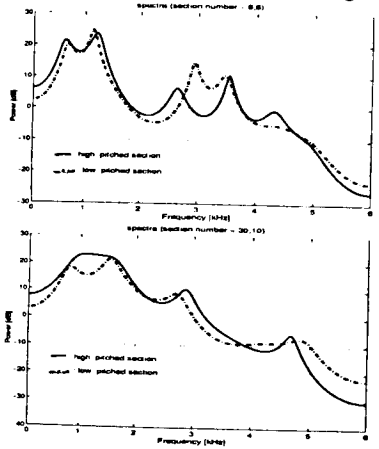


Fig.7 VTTF for /a/ in /aká/

In Fig.8, the estimated formant traces are shown for the second vowel /a/ in the VCV /áka/ (upper graph) and /aká/ (lower graph). The upper trace corresponds to the case of low pitch frequency, and the lower trace is high. From these experiments, we see that the pattern of the formant trajectories is different in the two graphs, and the formants are influenced much by the pitch frequency or the accent. The fluctuation of formant trajectories is larger for the low pitch (non-accent) than the high pitch (accent) vowel.

As our conclusion, since the pitch accent causes movement of the glottal position even in the same articulation effort for the vowel, the VTTF or formant can be shifted by accent.

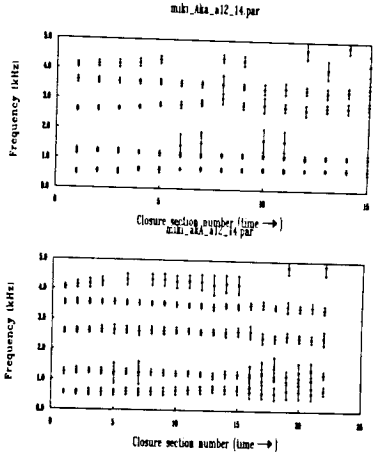


Fig.8 Traced formants for /a/ part in the second vowel in /áka/ (upper) and /aká/ (lower)

Acknowledgment We are grateful to Dr. Y. Pham Thi Ngoc, Dr. I. Karlsson, and Dr. C.H. Shadle for there help in our data aquisition.

References

[1] Miki N., Takemura K., and Nagai N., "A Short-time Speech Analysis Method with Mapping Using the Fejer Kernel," IEICE Trans. Fundamentals, E77-A, pp.792-799 (1994)

[2] Miki N., Badin P., Pham Thi Ngoc Y., and Ogawa Y. (1994)," Vocal Tract Model and 3-Dimensional Effect of Articulation," ICSLP94, 1,pp167-170.

[3] Hirai H. and Honda K., "Analysis of magnetic resonance images on the physiological mechanisms of fundamental frequency control," J. Acoust. Soc. Jpn. vol.50, 4,pp.296-304 (1994) (in Japanese)

[4] Badin P.(1991), "Fricative consonants: acoustic and X-ray measurements," J. Phonetics, 19, pp397-408

A COOPERATIVE APPROACH TO FORMANT EXTRACTION

Alain Soquet
Institut des Langues Vivantes et de Phonétique
Université Libre de Bruxelles

ABSTRACT

In this paper, we investigate a way to improve formant extraction reliability by using a cooperative approach. The basic motivation is that extraction methods based on different principles should not fail simultaneously when evaluating the same quantity. Therefore, we propose to combine independent formant extractors. Each method provides a set of formant candidates. These candidates are then combined with a vote mechanism in order to keep those that most probably correspond to a formant and to reject the majority of spurious values.

INTRODUCTION

Problems in formant frequency estimation are due both to the difficulty of estimating the resonances of the vocal tract at any given time (formant extraction), as well as to the difficulty of obtaining reasonable contours (formant tracking). Most formant tracking algorithms use the output of the extraction algorithm for successive segments and try to detect and correct extractor mistakes by using speech knowledge expressed, for example, by heuristic rules [1] or statistical models [2].

People are faced with many difficulties when trying to estimate the formant frequencies. A first problem is related to the coupling between the excitation and the vocal tract. For a given vocal tract configuration, the complexity of the estimation of the formant frequencies depends on the acoustic excitation and on the position where this excitation takes place; the difficulties encountered for high pitched voices are well known. A second problem is related to the precision with which the formant frequencies can be determined. Lindblom [3] estimated the accuracy of spectrographic measurement to be approximately equal to a quarter of the fundamental frequency. Monson et al. [4] compared the accuracy of spectrographic techniques and of linear prediction analysis in measuring formant frequencies on synthetic speech tokens. They observed that, "for fundamental frequencies between 100 and 300 Hz, both methods are accurate to within approximately ±60 Hz for both first and second formants. The third formant frequency can be measured with the same degree of accuracy by linear prediction, but only to within ±110 Hz by spectrographic means. The accuracy of both methods decreases greatly when fundamental frequency is 350 Hz or greater". This study clearly illustrates the degree of accuracy that can be expected from a given extractor.

In order to improve the general performances of the extraction, we propose to combine formant candidates provided by different basic extractors.

BASIC EXTRACTORS

The speech signal was passed through a 5 kHz cutoff low-pass filter, and sampled at 10kHz. The signal was then preamphasized ($1 - 0{,}95\ z^{-1}$) before further processing.

We have used three well documented basic extractors. We have chosen the linear prediction [1], the cepstrum [6] and the group delay functions [7].

- **Linear prediction** (LPC): The LPC coefficients were computed with the autocorrelation method on a 25,6 ms frame multiplied by a Hamming window. The number of poles of the predictive filter was fixed to 12. The formant frequencies can be estimated from these coefficients by different means (see Christensen et al. [5] for a discussion).
- **Cepstrum**: The second method is based on cepstral smoothing [6]. The cepstral coefficients were computed from a 16 ms frame multiplied by a Hamming window. The parameters of the cepstral filtering have been chosen as suggested in [6] and [8]. In order to enhance the formant resolution in the smoothed spectra, we used the Chirp-Z transform [6]: this transform consists in evaluating the spectrum on a circle of radius $\alpha < 1$.
- **Group Delay Function** (GDF): The group delay functions are the negative derivative of the Fourier transform phase. We used the method proposed in [7]. This method involves deriving a signal with the characteristics of a minimum phase signal. The peaks of the GDF derived from this phase function correspond to formants.

COOPERATIVE APPROACH

The basic motivation is that extraction methods based on different principles should not fail simultaneously when evaluating the same quantity.

We adopted a majority vote among M methods based on the following principle. Let $C^i(f)$ be the set of candidates provided by the i^{th} method. $C^i(f)$ is different from zero only for the values of f corresponding to a candidate of the i^{th} method. The procedure consists in six stages:

1. Let $F_{min} = 0$.
2. Search for the first candidate such as $f > F_{min}$. The procedure stops if no candidate is found.
3. Let $N = 0$.
4. For each method i, look between the frequencies f and $f + \Delta F$ (with ΔF the length of the search interval) for the candidate with the lowest frequency. If it does exist, put its frequency and amplitude respectively in f_i and c_i, and increment N. Otherwise let $f_i = 0$ and $c_i = 0$.
5. **If** $N \geq N_{min}$ **then** a candidate of frequency F is proposed by the cooperative approach:

$$F = \frac{\sum_{i=1}^{M} f_i c_i}{\sum_{i=1}^{M} c_i} \qquad (1)$$

and its amplitude $C^{vote}(F)$ is given by:

$$C^{vote}(F) = \sum_{i=1}^{M} c_i \qquad (2)$$

Let $F_{min} = F + \delta F$, with δF the minimal frequency difference between two successive formants.

If $N < N_{min}$ **then** let $F_{min} = f$.

6. Back to step 2.

Two iterations of this procedure are presented graphically in figure 1 for $M = 3$ and in the case of an unanimous vote ($N_{min} = 3$).

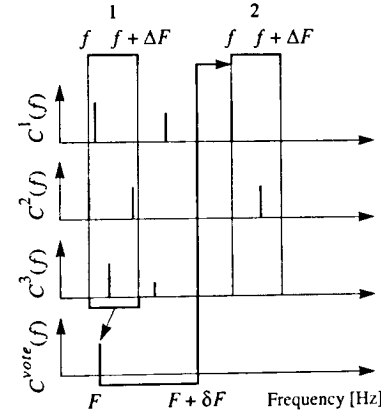


Figure 1: Example of two iterations of the combination mechanism for formant frequency estimation.

The parameters of the cooperative approach have been chosen as follows. The window length ΔF is related to the precision of the evaluation by the different methods; we fixed $\Delta F = 200Hz$. δF correspond to the minimal distance between two consecutive formant frequencies; we chose $\delta F = 100Hz$. The choice of N_{min} depends on the performances of the basic extractors: if the extractors tend to propose too many candidates with the formants among them, an unanimous vote could be a good choice. On the contrary, if the extractors tend to miss formants, a vote at the absolute majority could be more adequate. We will thus come back on the choice of this parameter during the evaluation.

We have chosen $M = 3$ and used the three basic extractors described above. The different set of candidates are

derived from the basic methods as follows:

- $C^{lpc}(f)$ is created by finding the poles of the LPC transfert function and by taking the corresponding amplitudes of the LPC spectrum.
- $C^{cepstre}(f)$ is obtained by picking peaks of the Chirp-Z transform and by taking the corresponding amplitudes of the smoothed spectrum.
- $C^{gdf}(f)$ is obtained by picking peaks of the group delay function.

EVALUATION

The basic extractors and the cooperative approach have been evaluated on a corpus of VCV logatomes, with V a vowel among [a, œ, i, u, y] and C a plosive among [p, t, k, b, d, g]. The corpus has been produced by three male speakers and segmented manually, in order to locate the segments with formantic structure. A measurement of the first four formants has then been made every 10 msec giving a total of 11385 measurements. The reference has been obtained manually on the basis of different representations of the speech signal. We focussed on two sets of rough errors: the insertion errors (see table 1) and the omission errors (see table 2).

Table 1: Notations used for insertion errors.

Location of the insertion	Notation
before F1	$\times F1$
between F1 and F2	$F1\times F2$
between F2 and F3	$F2\times F3$
between F3 and F4	$F3\times F4$

Table 2: Notations used for omission errors.

Omission of one formant	Notation
F1	$\not F1$
F2	$\not F2$
F3	$\not F3$
F4	$\not F4$

The experiment has been conducted for two versions of the cepstrum depending on the value of the Chirp-Z transform coefficient: $\alpha = 0.95$ and $\alpha = 0.8$.

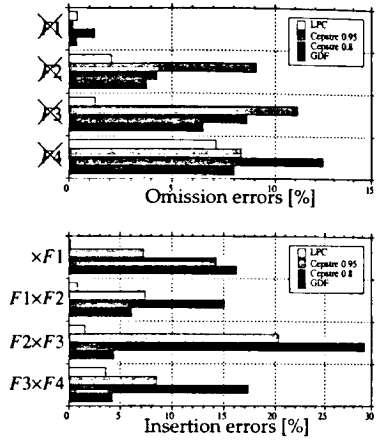


Figure 2: Omission and insertion errors for the different basic methods LPC, cepstrum ($\alpha = 0.95$ and $\alpha = 0.95$) and GDF.

The figure 2 shows the results for basic extractors: LPC, cepstrum ($\alpha = 0.95$ and $\alpha = 0.8$) and GDF.

It can be seen that LPC is the most reliable of the four methods, in every error category. GDF turns out to give very satisfactory results with reasonable error rates both for omissions and insertions (except for the category $\times F1$). For cepstrum, the lowering of the omission error rates related to the use of $\alpha = 0.8$ instead of $\alpha = 0.95$ is quite clear, but causes an important increase of the number of insertions. The reliability of this method remains quite low.

Given the important amount of insertion errors caused by cepstrum and GDF, we chose a unanimous vote ($N_{min} = 3$).

The method VOTE 1 is obtained by combining the candidates of the extractors LPC, cepstrum with $\alpha = 0.95$ and GDF. VOTE 2 is obtained by combining the candidates of the extractors LPC, cepstrum with $\alpha = 0.8$ and GDF.

The results for the individual extractors and the cooperative approach are presented on table 3 and table 4 respectively for the omission and the insertion errors.

It can be seen that the insertion error rates of the cooperative approaches are extremely low in comparison with the individual extractors. The lowering varies from a factor 10 with the LPC to 100

Table 3: Comparison of omission errors for LPC, cepstrum ($\alpha = 0.95$ et $\alpha = 0.8$), GDF, VOTE 1 and VOTE 2 on the whole corpus.

Method	$\not F1$	$\not F2$	$\not F3$	$\not F4$
LPC	46	235	146	810
Cepstre 0,95	21	1032	1261	949
Cepstre 0,8	142	487	982	1411
GDF	42	432	741	911
VOTE 1	76	1050	1220	1054
VOTE 2	170	589	878	1051

Table 4: Comparison of insertion errors for LPC, cepstrum ($\alpha = 0.95$ et $\alpha = 0.8$), GDF, VOTE 1 and VOTE 2 on the whole corpus.

Method	$\times F1$	$F1\times F2$	$F2\times F3$	$F3\times F4$
LPC	14	95	177	407
Cepstre 0,95	825	845	2315	970
Cepstre 0,8	1626	1712	3328	1974
GDF	1845	697	497	481
VOTE 1	6	4	22	40
VOTE 2	2	3	32	67

for cepstrum with $\alpha = 0.8$. This result clearly confirms the main hypothesis of the cooperative approach.

Unfortunately, the use of a unanimous vote has an important drawback: the omission error rates are comparable with those of the less performant extractor participating to the vote. The gain in performance of VOTE 2 compared to VOTE 1 directly reflect the lowering of omission errors of the cepstrum with $\alpha = 0.8$ instead of $\alpha = 0.95$.

CONCLUSION

The results show that the use of a cooperative approach allows the suppression of most of the insertion errors. Indeed, the number of insertion errors is reduced by a factor 10 to 100, depending on the individual method chosen as reference. This clearly illustrates the basic advantage of cooperative approach: the candidates proposed by the vote mechanism are likely to correspond to formant values. However, we have noted that the results for omission errors are comparable to those obtained by the least successful method used for the vote. Therefore, the individual methods have to be chosen so as to have a low omission error rate even if the insertion error rate is relatively high, since most of the insertion errors are eliminated by the vote mechanism.

ACKNOWLEDGMENTS

This work was partially supported by the "Communauté Française de Belgique" in the framework of the ARC 93/98 — 168 project.

REFERENCES

[1] J. D. Markel, and A. H. Gray, "Linear prediction of speech," Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1976.

[2] G. E. Kopec, "Formant tracking using hidden markov models and vector quantization," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 34, n°4, pages 709-729, 1986.

[3] B. Lindblom, "Accuracy and limitations of sonagraph measurements," Proceedings of the 4th International Congress of Phonetic Sciences, Helsinki,The Hague, 1962.

[4] R. B. Monsen, and A. M. Engebretson, "The accuracy of formant frequency measurements: a comparison of spectrographic analysis and linear prediction," Journal of Speech and Hearing Research, vol. 26, pages 89-97, 1983.

[5] R. L. Christensen, W. J. Strong, and E. P. Palmer, "A comparison of three methods of extracting resonance information from predictor-coefficient ceded speech," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, n°1, pages 8-14, 1976.

[6] R. W. Schafer, and L. R. Rabiner, "System for automatic formant analysis of voiced speech," J. Acoust. Soc. Am., vol. 47, n°2, pages 634-648, 1970.

[7] H. Murthy, and B. Yegnanarayana, "Formant extraction from Group Delay Function," Speech Communication, vol. 10, pages 209-221, 1991.

[8] Y. Laprie, "Formant tracking adapted to acoustic-phonetic decoding," Eurospeech, pages 669-672, 1989.

A METHOD FOR TRACING NASALITY

Mechtild Tronnier
Dept. of Linguistics and Phonetics,
Lund, Sweden

ABSTRACT

This paper aims to present a method of tracing nasality in speech with a very convenient device also accessible to researchers with only restricted lab facilities. The basic idea lies in using the spectral information represented by the formant pattern in a signal, recorded with a contact microphone attached to the nose.

INTRODUCTION

Due to its difficult accessibility, the description of the movements of the velum has demanded quite some imagination and technology to find appropriate methods for this purpose. Such methods vary from nasal airflow measurements, fiberoptic endoscope [1], velotrace [2], cineradiographic techniques [3] and X-ray microbeam [4] to the description of spectral events in the spectrogram of the speech signal [5, 6 & 7]. However, these methods are either invasive and/or demand valuable and expensive equipment to be effective. Furthermore, some of these devices do not permit an undisturbed simultaneous recording of the speech signal. The use of a contact microphone attached to the outer part of the nose has also been introduced [8]. With the help of such an accelerator microphone, the increase of vibration of the skin at the outer part of the nose, which is due to nasal airflow, can be measured. Interference from the oral signal, however, seems to be an obstacle which makes it difficult to determine whether a certain increase in amplitude of the signal recorded at the nose is related to nasal airflow or whether it originates from the oral signal. Making use of the intensity curve of the nasal signal has been its usual way of analysis: diverging directions of the aligned intensity curves of both the nasal and the oral signal were judged to be an indication of absence of any interference on the nasal signal by the oral signal.

In this paper an additional method will be presented, which makes use of the spectral information of the signal obtained with a contact microphone. Such a method does not give a physiological explanation about the exact behaviour of the velar movements and its relationship to other articulators, as do some of the devices mentioned above, but it aims to provide phoneticians and speech pathologists with a reasonably priced tool to detect onset and offset of nasality in undistorted speech. The analysis procedure of the nasal signal is related to the traditional - visual, auditory and manual - spectrogram analysis of the standard speech signal. Recorded French speech, containing CV, CṼ and CVN sequences, was used to evaluate this method.

PROCEDURE

With a two channel DAT-recorder, the oral and the nasal signal of French speech, spoken by a native male speaker, were recorded, using the contact microphone for the nasal signal. The recorded material contained sequentially read words of French, including CV, CṼ and CVN sequences. Some of those words were e.g. dos, dent, donne, etc.

During the recording procedure, the contact microphone was attached to the upper part of the nose on one side, where the lateral nasal cartilage is found (Figure 1). This location was discovered to be the most appropriate one for achieving a suitable nasal signal [9]. The contact microphone is a lightweight accelerometer named *Hot Spot* made by *K&K Sound Systems*. It consists of a small metallic disc which has a diameter of 12mm and is 0.7mm thick. It is attached to either the nose or - as originally intended - to acoustic musical instruments with an adhesive strip. The frequency response goes from 20-15000Hz. Unfortunately, the producer of the microphone did not provide us with more detailed information about the frequency response of the microphone. It has been reported though, that "any other accelerometer sensitive to vibrations in the frequency region 100-1500Hz could be used to detect nasalization" [9]. However, this microphone is also sensitive above the frequency region of 1500Hz. No preamplifying is needed with this microphone. Such a microphone costs about US$35.

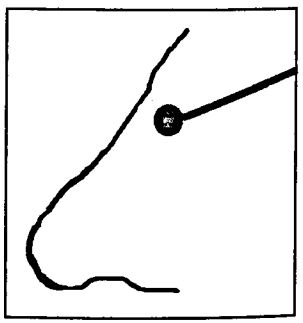


Figure 1. Placement of the contact microphone.

The signals of both channels and their spectra were analysed and displayed in the ESPS/Waves+ environment, and auditory and manual labelling was carried out.

OBSERVATIONS

Figures 2 to 4 present the French words *dent* , *dos* and *donne* in five windows, where the upper one shows the waveform of the nasal channel recorded with the accelerator microphone. In the next window below, the spectrogram of the nasal waveform is displayed. The third window shows the waveform of the general speech signal and its spectrogram follows in the window underneath. In the bottom window, the labels are marked. Here, the vowels should be read as follows, since the ESPS/Waves+ tool does not provide any IPA-font: |o| reads as [o], |e| as [a], |O| as [õ] and |E| reads as [ã].

In the second window of Figure 2, which presents the spectrogram of the nasal signal of the word *dent*, formant structure for the nasal vowel [ã] can be observed. However, the nasalization - presented by such a formant pattern - does not start right from the beginning of the vowel after the release of the voiced stop [d]. This is also audible, when listening to the speech signal. Figure 3, which presents the French word *dos* , does not show any formant structure in the second window at all. This utterance does not contain any nasal vowel or consonant. In Figure 4, however, presenting the word *donne*, formant structure can be observed in the second

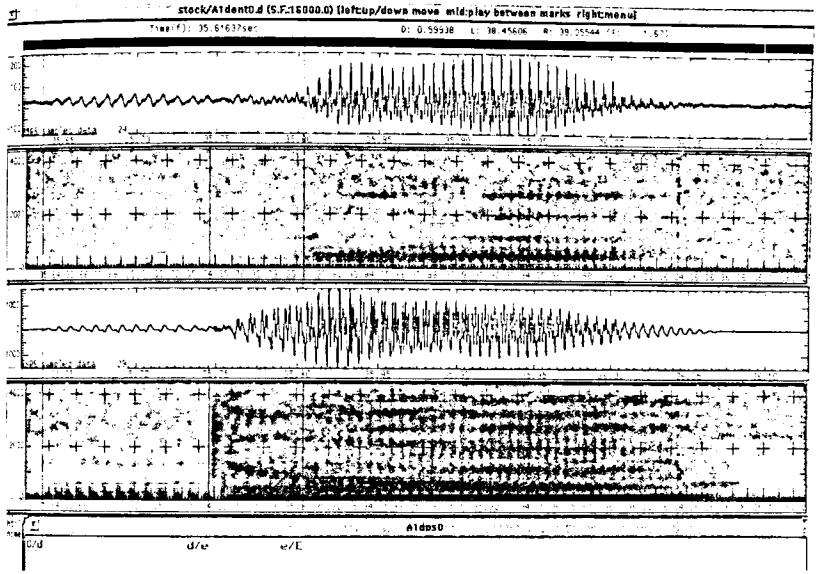


Figure 2. Nasal signal and speech signal representation of the waveforms and the spectrograms and the phonetic labels of the French word dent, in IPA-transcription [dã].

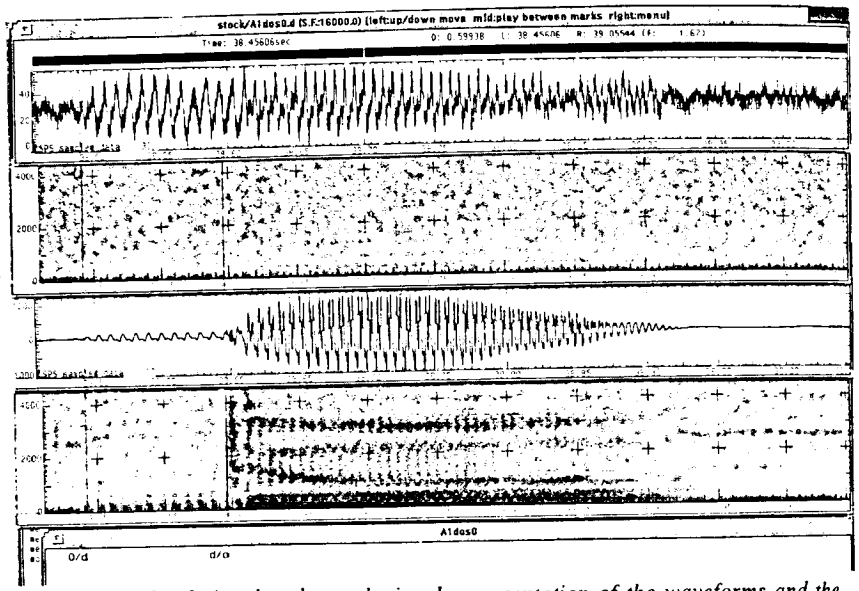


Figure 3. Nasal signal and speech signal representation of the waveforms and the spectrograms and the phonetic labels of the French word dos, in IPA-transcription [do].

window not only during the production of the nasal consonant [n], but also at the end of the phonologically non-nasalized vowel /o/ preceding the nasal consonant. When listening to the speech signal, this portion of the vowel clearly can be heard as being nasalized. Very faint formant structure can be found in the vowel already earlier on (labelled *N*), which also does not start until only later after the release of the voiced stop [d]. This kind of nasal activity is not observable when listening to the speech signal. It could however reflect anticipatory activity of the velum being lowered. Although using the same intensity/grey relationship for the display of the spectrogram of the nasal signal across the utterances, even such faint formant structure does not show up in the complete non-nasal environment (Figure 3).

The use of the spectral information of the nasal signal has its advantages over the use of the intensity curve or the waveform only. Even though one should notice that the amplitude scaling for the nasal signal varies across the utterances, one could be mislead in that periodicity in the waveform with greater amplitude would reflect a lowered velum positioning and therefore a certain degree of nasality. It has been reported in earlier work, that interference with the fundamental frequency and the signal obtained with the accelerometer microphone at the nose is most likely to occur. Using the spectral information of a nasal signal could thus extract such obstacles.

As presented earlier [10], the values of the formants observed in the spectrogram of the nasal signal correspond to the formant values found in the speech signal. The second formant shows a lower bandwidth which denotes a sharper peak correlated with higher ener-gy for the signal obtained at the nose in contrast to the usual speech signal.

CONCLUDING REMARKS

A method of measuring nasality was suggested above by making use of the spectrogram of a signal obtained with a contact microphone attached to the nose. Clear formant structure is visible here in the case of the production of nasal consonants, nasal vowels and at a certain period of time during the production of phonologically non-nasal vowels in the vicinity of nasals. In analogy to the traditional way of spectral analysis this low cost tool is a very convenient instrument for phoneticians and speech pathologists who have only limited lab facilities, but have access to spectral analysis tools. For the sake of orientation and alignment, a two channel display showing the nasal signal and the speech signal simultaneously would be preferable.

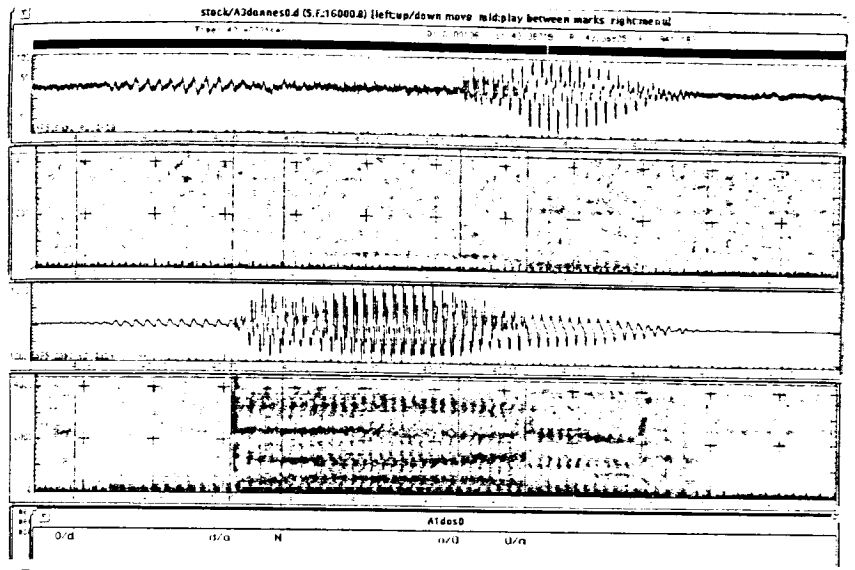


Figure 4. Nasal signal and speech signal representation of the waveforms and the spectrograms and the phonetic labels of the French word donne, in IPA-transcription [dɔn].

ACKNOWLEDGEMENTS

The author would like to thank the speaker Paul Touati.

[1] Ushijima, T. & Sawashima, M. (1972), "Fiberscopic observations of velar movement during speech", *Ann. Bull. RILP* 6, pp. 25-38.

[2] Niimi, S. & Bell-Berti, F. (1987), "The Velotrace: A Device for Monitoring Velar Position", *Cleft Palate Journal*, 24 pp. 104-111.

[3] Kuehn, D.P.(1976),"A cineradiographic investigation of velar movement variables in two normals", *Cleft Palate Journal*, 13, pp. 88-103.

[4] Kiritani, S., Itoh, K., & Fujimura, O. (1975), " Tongue-pellet tracking by a computer controlled X-ray microbeam system", *J. Acoust. Soc. Am.*, 57, pp. 1516-1520.

[5] Fujimura, O. (1976), "Analysis of nasal consonants", *J. Acoust. Soc. Am.*, 34, pp. 1865-1975.

[6] Hattori, S., Yamamoto, K. & Fujimura, O. (1958), " Nasalization of Vowels in Relation to Nasals", *J. Acoust. Soc. Am.*, 30, pp. 267-274.

[7] Schwartz, M.F. (1971) "Acoustic Measures of Nasalization and Nasality", in Grabb, Rosenstein & Bloch (eds): *Cleft, Lip and Palate : Dental and Speech Aspects*, pp. 798-804

[8] Stevens, K.N., Kalikow, D.N. & Willemain, T.R. (1975) "A miniature accelerometer for detecting glottal waveforms and nasalization", *J. Speech and Hear. Res.*, 18, pp. 594-599.

[9] Lippmann, R.F. (1981) "Detecting nasalization using a low-cost miniature accelerometer", *J. Speech and Hear. Res.* 24, pp. 314-317.

[10] Tronnier, M. (1994) "Tracing Nasality with the help of the Spectrum of a Nasal Signal", *Proceedings of the Fifth Australian Conference on Speech Science and Technology*, Perth, pp. 330-335.

VARIATIONAL METHOD APPLIED TO FORMANTS COMPUTATION FOR A PHARYNGO-BUCO-NASAL TRACT

R. Van Praag and P. Jospa
Inst. of Modern Languages and Phonetics , CP 110, Universite Libre de Bruxelles, 50 Avenue F.-D. Roosevelt, 1050 Bruxelles, Belgium.

Abstract

In this paper, we show how the variational method can be used in order to compute the formants and the sensitivity functions related to a vocal tract when the velum is open. The stationary modes of vibration obtain by this method (whithout the need to compute the transfer function) lead to a physical description of anti-formants.

1 Introduction

The variational formulation of the acoustico-articulatory link has already been successfully applied to a single vocal tract(disregarding the nasal tract)[1]. It provides us with a quick numerical algorithm allowing us to compute stationary vibration frequencies(the formants) for a given vocal tract whose geometry is represented by an area function.

This method may be extended to the case where the nasal tract is connected by imposing both flow conservation and pressure continuity at the velum.

2 Acoustic model

2.1 The Sondhi Model

We have adopted the acoustic modeling of the vocal tract proposed by Sondhi[2]. This model is characterized by a parietal admittance proportional to the area function, a constant shape factor, and a sound propagation in quasi planar wave fronts.

Let us consider the velocity potential $\Phi(x,t)$ where x is the distance from the glottis and t is the time. The volume velocity ν and the average pressure p are given by:

$$\nu(x,t) = -\Phi_{,x} \; ; \; p(x,t) = \rho\Phi_{,t} \qquad (1)$$

Henceforth,we will note either $\partial_x g$ or $g_{,x}$ for the partial derivative $\frac{\partial g}{\partial x}$ of function g. For a normal mode at frequency $f = \omega/2\pi$, the velocity potential may be written:

$$\Phi(x,t) = \Psi(x)e^{-\sigma t}cos(\omega t), \qquad (2)$$

where Ψ is the spatially distributed mode amplitude and σ represents the damping of the field caused by wall elasticity. In such a case, it can easily be demonstrated that $\Psi(x)$ verifies the following Webster equation:

$$\partial_x(A(x)\partial_x\Psi(x)) + \frac{(\omega^2-\omega_p^2)}{c^2}A(x)\Psi(x)=0 \qquad (3)$$

where ω_p is the resonance frequency of walls ($\sim 200 \times 2\pi s^{-1}$), and c, the sound velocity ($c = 34000cm/s$). For homogenous boundary conditions, this equation has solutions only for discrete values of ω.

2.2 Tract ends conditions

In this paper, we will consider the case where the glottis is closed and assume an infinite glottis impedance. Hence, we have:

$$\partial_x\Psi(0) = 0. \qquad (4)$$

Considering that the lip radiation can be approached by that of a vibrating piston set in a spherical baffle(the head) [1] leads us to :

$$A(L)\Psi_{,x}(L) + q\sqrt{A(L)}\Psi(L) = 0 \qquad (5)$$

with L being the total tract length(i.e the distance between the glottis and the lips), and q a factor depending on the degree of aperture of the lips. As a first empirical estimation, we propose:

$$q = a\sqrt{A(L)} + b \qquad (6)$$

with $a = -3.5cm^{-1}$ and $b = 35.0$

2.3 Three tracts linked together

In this case, we will need three velocity potential functions in order to describe the wave behavior in a three tracts system. Let Ψ_{ph},Ψ_b and Ψ_n be those functions for respectively pharyngal, bucal and nasal tracts.

For the sake of simplicity we change the variable x into z such as:

$$z_\alpha = x_\alpha/L_\alpha \qquad (7)$$

with α taken as ph, b or n. We have thus, for each tube the related Webster equation (4) which is now written:

$$\partial_z(A_\alpha\partial_z\Psi_\alpha)+((\omega^2-\omega_{p\_\alpha}^2)\frac{L_\alpha^2}{c^2})A_\alpha\Psi_\alpha=0 \qquad (8)$$

We emphasize in eq (8) that ω doesn't take an symbol α (ph, b or n),it represents the angular resonance frequency of the whole three tubes system.These are coupled together by imposing both pressure continuity:

$$\Psi_{ph}(1) = \Psi_b(0) = \Psi_n(0) \qquad (9)$$

and flow conservation:

$$\frac{A_{ph}(1)}{L_{ph}}\partial_z\Psi_{ph}(1) - \frac{A_b(0)}{L_b}\partial_z\Psi_b(0) - \frac{A_n(0)}{L_n}\partial_z\Psi_n(0) = 0 \; . \qquad (10)$$

The glottis condition (4) is

$$\partial_z\Psi_{ph}(0) = 0, \qquad (11)$$

and the lip and nostril conditions (5) become (by 7):

$$A_b(L_b)\partial_z\Psi_b(L_b)+q\,L_b\sqrt{A_b(L_b)}=0 \qquad (12.1)$$

$$A_b(L_n)\partial_z\Psi_n(L_n)+q\,L_n\sqrt{A_n(L_n)}=0 \qquad (12.2)$$

3 Variational method

Let

$$I=\int_{z_0}^{z_1}\mathcal{L}(z,\Psi(z),\Psi'(z))dz + G(\Psi(z_0),\Psi(z_1)) \qquad (13)$$

The extremals of I (here, minimal) for which $\delta I = 0 \; \forall\delta\Psi \ll \Psi$ verify the system:[3]

$$\frac{\partial\mathcal{L}}{\partial\Psi} - \frac{d}{dz}\frac{\partial\mathcal{L}}{\partial\Psi'} = 0 \qquad (14.1)$$

$$\frac{\partial\mathcal{L}}{\partial\Psi'}|_{z_1} + \frac{\partial G}{\partial\Psi(z_1)} = 0 \qquad (14.2)$$

$$\frac{\partial\mathcal{L}}{\partial\Psi'}|_{z_0} - \frac{\partial G}{\partial\Psi(z_0)} = 0 \qquad (14.3)$$

We will now build three functionals(I_α, $\alpha = ph, b, n$) whose

extremals are solutions of Webster equations (8) with respect to the junction conditions (9) and (10) and the boundaries conditions (11) and (12).

Let

$$\mathcal{L}_\alpha(z,\Psi_\alpha,\Psi'_\alpha)=\frac{1}{L_\alpha}A_\alpha(z)((\partial_z\Psi_\alpha(z))^2 - \frac{L_\alpha^2}{c^2}(\omega^2-\omega_{p,\alpha}^2)\Psi_\alpha(z)^2) \quad (15)$$

α taken as ph, b or n and

$$G_{ph}=-2\Psi_{ph}(1)[\frac{A_b(0)}{L_b}\partial_z\Psi_b(0)+\frac{A_n(0)}{L_n}\partial_z\Psi_n(0)] \quad (16)$$

$$G_b=2\Psi_b(0)[\frac{A_{ph}(1)}{L_{ph}}\partial_z\Psi_{ph}(1)-\frac{A_n(0)}{L_n}\partial_z\Psi_n(0)] + \Psi_b(0)[2\Psi_{ph}(1)-\Psi_b(0)]+q\sqrt{A_b(1)}L_b\Psi_b^2(1)$$

($G_n = G_b$ transposing b and n) be the $\mathcal{L}$ and G functions for the three functionals. The extremals of those functionals I_α are solutions of the following variational system:

$$\delta I_\alpha/\delta\Psi_\alpha = 0, (\alpha = ph, b, n) \quad (17)$$

and thus verify the system (14).

Now,it is easy to show that the equation (14.1)→(14.3) applied to (15) & (16) gives us the wanted Webster equations (8), pressure continuity (9), flow conservation (10) and the glottis, lip and nostril conditions (11) and (12).

4 Resonance mode computation

The extension of the Rayleigh-Ritz[4] method in the case where three functionals have to be minimized together will lead to a fast and precise algorithm allowing us to compute resonance modes frequencies and the associated wave functions .

Let η_i be a set of n linearly independent functions(we have chosen here Tchebychev polynoms). Suppose that three functions Ψ_{ph},Ψ_b and Ψ_n verify (17) and so are solutions of equations (8)→(12).These functions may be written as an approximation such as

$$\begin{aligned}\Psi_{ph}(z) &= \textstyle\sum_{i=1}^{n} C_i\eta_i(z)\,;\\ \Psi_b(z) &= \textstyle\sum_{i=n+1}^{2n} C_i\eta_{i-n}(z)\\ \Psi_n(z) &= \textstyle\sum_{i=2n+1}^{3n} C_i\eta_{i-2n}(z)\end{aligned} \quad (18)$$

where,here for the sake of simplicity, we have taken the same number of approximation functions for each of the three tubes.

Injecting (18) into (16) we can write

$$I_\alpha = I_\alpha(C_1,...,C_{3n},\eta_1,...,\eta_n,\eta_1',...,\eta_n') \quad (19)$$

for α as ph,b or n.

Looking for the extremals of these three functionals,we will impose:

$$\begin{aligned}\frac{\partial I_{ph}}{\partial C_i} &= 0 \quad \forall i \in (1,...,n)\\ \frac{\partial I_b}{\partial C_i} &= 0 \quad \forall i \in (n+1,...,2n)\\ \frac{\partial I_n}{\partial C_i} &= 0 \quad \forall i \in (2n+1,...,3n)\end{aligned} \quad (20)$$

after having replaced the Ψ_α by their developments (18).

Thus we obtain three systems of n equations and $3n$ unknowns coupled together thanks to the junction terms and giving one system of $3n$ equations and $3n$ unknowns of the form

$$\sum_{i=1}^{3n}(V_{ij}-\omega^2 W_{ij})C_j = 0 \quad (21)$$

for $i \in (1,..,3n)$ which can be written in matrix form.

Multiplying on the left by W_{ij}^{-1},we obtain

$$(W^{-1}V - \omega^2\mathbb{1})\overline{C} = \overline{0} \quad (22)$$

$\mathbb{1}$ being the unitary matrix $\mathbb{R}^{3n}\rightarrow\mathbb{R}^{3n}$.

(22) has non trivial solutions if and only if ω^2 is eigen value of the operator $W^{-1}V$. So the first s $(\omega_1,...,\omega_s)$ eigenvalues(associated to the formants) and corresponding eigen vectors (the $\overline{C}_s$ giving the solutions (18)) can be computed easily by classical numerical methods.

Given the stationarity of the functional for the l^{th} mode $\Psi_{\alpha,l}$, the sensitivity functions can be obtained from the conditions:

$$I_\alpha(\omega_l+\delta\omega_l, A(x)+\delta A(x), \Psi_{\alpha,l}, \partial_x\Psi_{\alpha,l}) = I_\alpha(\omega_l, A(x), \Psi_{\alpha,l}, \partial_x\Psi_{\alpha,l}) \quad (23)$$

where $\delta A(x)$is an area function perturbation localised at x. These conditions are correct at the first order of perturbations.

4.1 Results

We will briefly discuss our results using a simple geometrical configuration. Considering a system of three uniform tracts, such as:
$L_{ph} = L_b = 8.5cm, L_n = 9cm$,
$A_{ph}(0) = 0, A_{ph}(x) = 5cm^2$ and $A_b(x) = A_n(x) = 2.5cm^2$, fig 1 shows two stationary modes extracted from the set obtained.

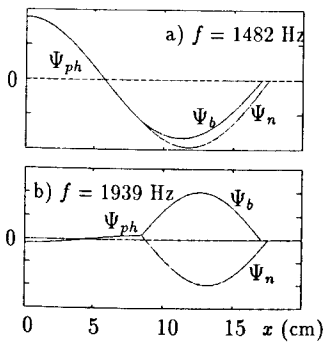


fig 1:wave function amplitude for two stationary modes

For a given resonance frequency, the pressure as function of time must be seen as the the projection of wave function Ψ rotating at angular frequency $\omega = 2\pi f$ around the x axis. The first (fig 1.a) shows a passing resonance,the acoustic impulsion($A\partial_x\Psi$) outgoing from pharynx being shared between the nasal and bucal tracts.

In fig 1.b , both potential($\div\Psi^2$) and kinetic($\div\partial_x\Psi^2$) energy are low at the end of the pharynx. No significant signal will cross the junction if the source is at the glottis. We think that such a normal mode will lead to an anti-formant and we intend to test our predictions on an experimental device. We can already say that, for some realistic geometrical configurations, we observe both a lowering of the first formant F1 and the occurance of a nasal formant between F1 and F2 which are characteristics of nasalised vowels described in the literature[5].

References

[1] P.Jospa, A.Soquet & M.Saerens (1995) Variational formulation of the acoustico-articulatory link and the inverse mapping by means of a neural network. *Levels in speech communication.* Elsevier, Amsterdam:103-113.

[2] M.Sondhi(1974): Model for wave propagation in a lossy vocal tract. *J. Acoust.Soc.Am,* 57(5):1070-1075.

[3] S.H Gould(1966): *Variational Methods for Eigenvalue Problems.* Oxford University Press.

[4] R.Courant & D.Hilbert(1953): *Methods of mathematical physics.* Interscience Publ., New York,Vol. 1.

[5] S.Maeda(1993): Acoustic of vowel nasalisation and articulatory shifts in french nasal vowels *Phonetics and Phonology,vol. 5. Nasals Nasalisation, and the Velum.* :147-167.

ANALYSIS OF FRICATIVES USING MULTIPLE CENTRES OF GRAVITY

Alan A. Wrench
CSTR, University of Edinburgh, Edinburgh EH1 1HN, UK.

ABSTRACT

An algorithm for describing speech spectra in terms of multiple centres of gravity is compared to traditional methods for parameterising fricative spectra. LPC peak-picking analysis and single centre of gravity measures are compared with Multiple Centroid Analysis (MCA) and the strengths and weaknesses of this newer approach are discussed.

INTRODUCTION

Traditional Spectral Parameters

It has been known for many years that fricative spectra exhibit formants and anti-formants. It seems likely from the wealth of research done on vowel perception that formants play an important part in our perception of fricative quality and accordingly, fricative formant frequencies have been studied in an effort to correlate them with articulatory parameters. Formant estimates have been recorded manually by visually peak-picking the fricative spectrum [1][2]. Despite being time consuming and subjective, hand labelling Most researchers have held back from using automatic formant analysis procedures for parameterising fricatives. Heinz and Stevens [3] fitted pole-zero model by hand to spectra. However, it is well known that automatic estimation of pole-zero parameters cannot be solved directly and while there do exist useful automatic procedures [4] for estimating these parameters they have tended to be ignored by researchers in this field most likely because the methods are not implemented in speech analysis packages. Although the fricative spectrum contains zeros associated with the source constriction or cavities posterior to a source of frication, several researchers have used LPC (which assumes an all-pole model) to approximate the spectral contour and have either identified peaks by eye or applied automatic peak picking in order to identify the fricative formant structure [5]. Due to the varying number of formants (and anti-formants) it is not possible to identify an ideal order for the LPC analysis of voiceless fricatives. A low order can produce biased or unresolved formant estimates and poor estimation of bandwidth. A higher order reduces this bias but is susceptible to producing spurious peaks.

The lack of theoretical basis for using an all-pole model combined with practical problems outlined above has lead many researchers to use the simpler calculation of the centre of gravity as a "general property detector" [6] for parameterising fricative and other speech spectra. The underlying assumption of this approach to fricative analysis is that the spectrum can be modelled as a single normal distribution which may reflect the dominant front cavity formant. The mean (1st moment) of this distribution is termed the centroid and popularly referred to as the centre of gravity. Despite this model being underspecified for fricatives which exhibit two or more formants, the centre of gravity has often been used to measure relative changes between fricative productions [7][8]. By extending the parameters to include higher moments of this estimated normal distribution (corresponding to skewness and kurtosis), the ability of this model to discriminate between fricatives is enhanced [9][10].

A natural extension to this model to cope with fricative spectra containing more than one formant was proposed by Jassem [11]. To do this the spectrum was split into two or three partitions. He investigated several criteria for automatically determining the partitions but found that fixed partitions were sufficient. The purpose, however, was not to relate each centre of gravity to a formant but to use them as an abstract set of parameters which might be fed into a statistical classifier in order to distinguish between fricative phonemes.

Overview of MCA

Aware of benefits of centroid analysis to the estimation of fricative spectra, Crowe looked for a method of optimally fitting multiple centroids to a speech spectrum intending it as a means of estimating vowel formants. He successfully generalised the centre of gravity calculation using a global least-squares error criterion to determine the optimal partitioning of the spectrum and thus provided a principled method for determining multiple centroids of a single multi-modal spectral distribution [12].

In broad terms the algorithm works as follows. Taking, as an example, dual centroid analysis: It operates by evaluating centroids for every possible partitioning the spectrum into two and choosing the pair of centroids that result in best overall fit. For each possible position of the boundary, the centre of gravity is calculated for the part of the spectrum lying in each partition.

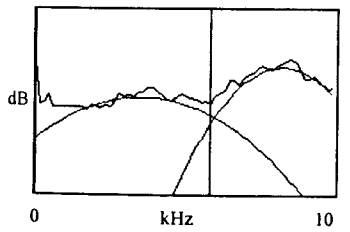


Figure 1. Optimum boundary indicated by vertical line and centres of gravity calculated for each partition (shown as the parabolic apex).

Each centroid is estimated as the frequency that gives the minimum squared error value. The two minimum error values are summed and stored for each possible boundary position. This process is repeated for all possible partitionings. The output of the analysis is the centroid pair corresponding to the partitioning with the lowest minimum error score.

The centroid can be thought of as fitting a Gaussian (normal) distribution to the power spectral distribution. The variance of this normal distribution can be thought of as an estimate of bandwidth and if the spectral distribution within a single partition contains a single formant then the centroid and associated variance represent the formant frequency and bandwidth. Multiple centroid analysis can be achieved more efficiently by placing constraints on how the spectrum may be partitioned and by bark scaling the spectrum prior to analysis. These measures improve the speed and accuracy of formant estimation when MCA is applied to vowel analysis [13].

Approach of the study

In this study we will take sustained examples of three voiceless fricatives and compare the results of analysing the speech using:

i) LPC analysis
ii) Single centre of gravity
iii) Multiple Centroid Analysis

Voiceless obstruents were chosen because they are known to be differentiated by their spectra alone.

The three speech segments were selected to contrast two allophones of /s/ (lip-rounded and non lip-rounded) with a non lip-rounded allophone of /ʃ/. These examples provide three distinct spectra which highlight the different behaviour of the three analysis methods.

METHODOLOGY

Recordings were made by a male speaker in an office environment using a Shure SM10A close-talking microphone and 16-bit soundcard sampling at 20kHz. Background noise levels were measured to be at least 25dB below the signal at all frequencies above 1khz.

The frame size used for all analyses was 6.4ms with a shift of 2 ms. In order to reduce the spectral variance, which can obscure the underlying resonant structure, a 100ms period of analysis was used. In the case of LPC analysis, the autocorrelation function was accumulated for each frame and averaged over a 100ms segment taken from the centre of the fricative. The LPC coefficients and associated spectrum were then calculated from this averaged function. Both the centre of gravity and MCA were based on time averaged spectra. An FFT was performed on each frame over the same 100ms portion of the fricative and the resulting power spectra were accumulated and averaged. The centre of gravity measure and MCA were calculated from the averaged power

spectral distribution lying above 1000Hz. Note that the MCA algorithm was applied directly to this spectrum and Bark scaling was not employed.

RESULTS

Figure 2a), b) and c) show the LPC (order 4) and FFT log spectra for /ʃ/, /ʂ/, and /ʂ/ respectively, extracted from the nonsense words 'eeshee', 'eessee', and 'oossoo'. The centre of gravity is calculated from the linear power spectrum. In order to represent the 1st and 2nd moments the centre of gravity is shown as the log of the corresponding normal distribution. Figure 3a), b) and c) show the same FFT spectra with LPC (order 6) superimposed and two centres of gravityestimated using MCA (order 2). Figure 4 a), b) and c) show the same FFT spectra with LPC (order 8) and three centres of gravity.

Using the general rule of thumb that two poles are required for each peak with two extra poles to model the spectral gradient, the order of the LPC analysis was selected to try to equate the maximum number of peaks with the number of centroids shown in the same figure.

DISCUSSION

We can see that the two-centroid analysis in Figures 3b and 3c model the formant structure well. The peaks of the LPC spectrum by visual comparison are biased. Increasing the order of the LPC analysis from 6 to 8 redresses this discrepancy in performance. In figure 3a the lower centroid matches the principal resonance well but, with no clearly defined second formant the upper centroid does not correspond to any feature. LPC, by contrast, models the spectrum in figure 3a by a single peak.

In formant tracking of vowels it is generally true that a fixed number of formants will exist within a given frequency range and having a fixed number of centroids is an advantage in that it permits merged formants to be resolved. In the case of fricative analysis, where the number of peaks varies as the length and shape of the cavities from source to lips changes, this advantage turns to disadvantage. This is clearly demonstrated in figure 4. where in each spectrum two centroids are associated with a single peak.

It is possible that, as abstract acoustic parameters, the 1st and 2nd moments of a pair of centroids may correlate well

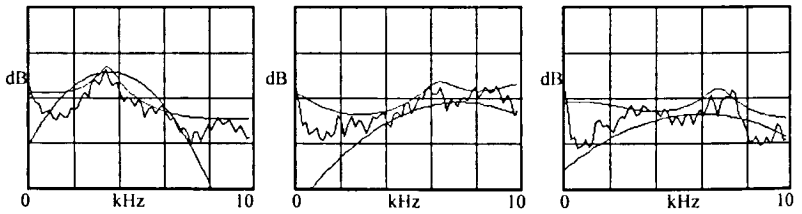


Figure 2. a) /ʃ/, b) /ʂ/ and c) rounded /ʂ/ 100ms segments analysed by LPC (order 4) and MCA (order 1) i.e. a single centre of gravity

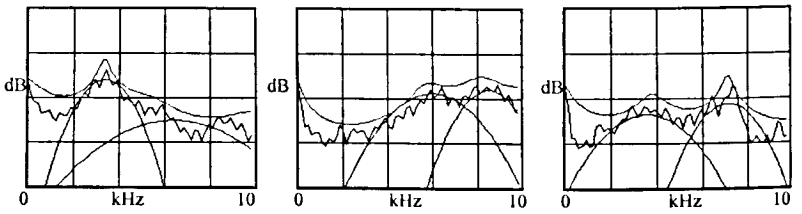


Figure 3. a) /ʃ/, b) /ʂ/ and c) rounded /ʂ/ 100ms segments analysed by LPC (order 6) and MCA (order 2) ie. optimised fit of two centres of gravity.

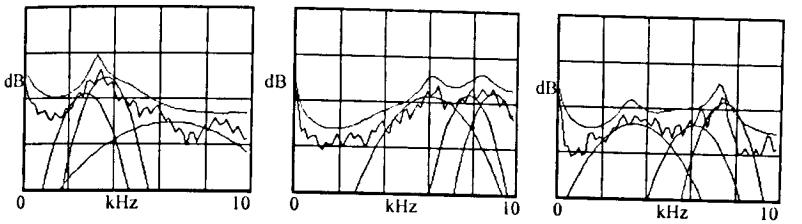


Figure 4. a) /ʃ/, b) /ʂ/ and c) rounded /ʂ/ 100ms segments analysed by LPC (order 8) and MCA (order 3) i.e. optimised fit of three centres of gravity

with articulatory parameters just as they have done for single centroid [9]. A rigourous comparative study is required to determine whether such parameters hold any advantage over the first 4 spectral moments of a traditional single distribution model or alternatively the set of centroids provided by fixed partitions as advocated by Jassem.

CONCLUSION

Unless MCA can be modified so that it automatically identifies the optimum number of centroids as well as their position, it is less useful than peak-picking LPC for the purpose of automatically identifying fricative formant structure. Multiple centroids may, however, be suitable as abstract parameters for fricative identification.

REFERENCES

1] Nguyen, N. and Hoole, P. (1993) Frequency variations of the lowest main spectral peak in sibilant clusters., *Proc. Eurospeech 93,* Vol. 1, pp 81-84.

[2] Mcgowan, R.S. and Nittrouer, S. (1988) Differences in fricative production between children and adults: Evidence from an acoustic analysis of /sh/ and /s/., *J. Acoust. Soc. Am.*, Vol 83, pp 229-236.

[3]Heinz, J.M. and Stevens, K.N. (1961) On the properties of voiceless fricative consonants, *J. Acoust. Soc. Am.*, Vol. 33 No. 5, pp 589-596

[4] Stieglitz, K. (1977) pp On the simultaneous Estimation of poles and zeros in speech analysis, IEEE Trans. ASSP, Vol. 25, No. 3, pp 223-233

[5] Soli, S.D. (1981) Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation, *J. Acoust. Soc. Am.* Vol. 70, pp. 976-984.

[6]Zue, V.W. (1991) From signals to symbols to meaning: On machine understanding of spoken language, Proc. Int. Conf. Phonetics, Speech, Vol. 1, pp 74-83

[7] Baum, S.R. and McNutt, J.C. (1990) An acoustic analysis of frontal misarticulation of /s/ in children. *Journal of Phonetics*, Vol. 18, pp 51-63.

[8] Nittrouer, S., Studdert-kennedy, M. and McGowan, R.S. (1989) The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal Speech and Hearing Research*, vol. 32, pp. 120-132.

[9] Forrest, K., Weismer, G., Milenkovic, P. and Dougall, R.N. (1988) Statistical analysis of word-initial voiceless obstruents: Preliminary data. *J. Acoust. Soc. Am.*, vol. 84, pp. 115-123.

[10] Nittrouer, S. (1995) Children learn separate aspects of speech production at different rates: Evidence from spectral moments. *J. Acoust. Soc. Am.*, Vol. 97, pp. 520-530.

[11] Jassem, W. (1979) Classification of fricative spectra using statistical discriminant functions. In Lindblom, B. and Ohman, S. (eds.) *Frontiers of Speech Communication Research*, (London: Academic Press), pp. 77-91.

[12]Crowe, A. and Jack, M.A. (1987) Globally optimising formant tracker using generalised centroids, *Electronic Letters*, Vol. 23, No. 19, pp 1019-1020

[13] Wrench, A.A., Watson, J.M.M., Soutar, D.S., Robertson, A.G., And Laver, J. (1994) Fast formant estimation of childrens speech. *Proc ICSLP94*, Vol. 3, pp. 1651-1654.

TRADING RELATIONS BETWEEN CUES FOR THE PHARYNGEALIZED/ NON PHARYNGEALIZED CONTRAST

Mohamed YEOU
Institut de phonétique, Sorbonne Nouvelle. CNRS , Paris.

ABSTRACT

The perceptual effects of orthogonal variations in two acoustic parameters (F1 and F2 onset frequencies) which differentiate Arabic pharyngealized /sˤ/ from plain /s/ were examined. An identification task showed a systematic displacement of the perceptual boundary as the onset value of F1 ($F1_0$) changes from low (250 Hz) to high (460 Hz), thus reflecting a trading relation between the two cues ($F1_0$ and $F2_0$). To investigate whether or not discrimination accuracy was differentially affected by the phonetic cooperation or conflict between the two cues, an AX discrimination task was used. As predicted, the discriminability ordering was the following: cooperating cues > one-cue > conflicting cues.

INTRODUCTION

In Arabic the four consonants /ð, t, d, s/ have the corresponding following pharyngealized consonants /ðˤ, tˤ, dˤ, sˤ/. These latter have in addition to a primary articulation (dental/alveolar contact), which they share with the former, a secondary articulation (backing of the tongue towards the pharyngeal wall). The acoustic consequence of this double articulation is a considerable lowering of F2 and a slight raising of F1 in vowels adjacent to pharyngealized consonants. In [1] locus equations which encode the dynamics of the F2 transition were capable of distinguishing pharyngealized consonants from non-pharyngealized ones. The purpose of this study is to investigate which acoustic properties identify pharyngealized consonants.

2. INDENTIFICATION TASK

2.1. Method

2.1.1. Stimuli

Four series of [s-sˤ] continua, each in the vowel context of [i:] were generated using a software parallel synthesizer [2]. Formant frequency values and timing characteristics for the [si:]-[sˤi:] series were adapted from the average values for a male Moroccan native speaker of Arabic. Two reference stimuli were used in the experiments: a pharyngealized type and a non-pharyngealized type. For the first type, F1 and F2 onset frequencies were 460 Hz and 1060 Hz, respectively. For the second type, the onset frequencies were 250 Hz and 1800 Hz. As Figure 1 shows the continua were constructed by systematically varying these onset frequecies in 10 steps of 35 Hz for F1 and in 10 steps of 140 Hz for F2.

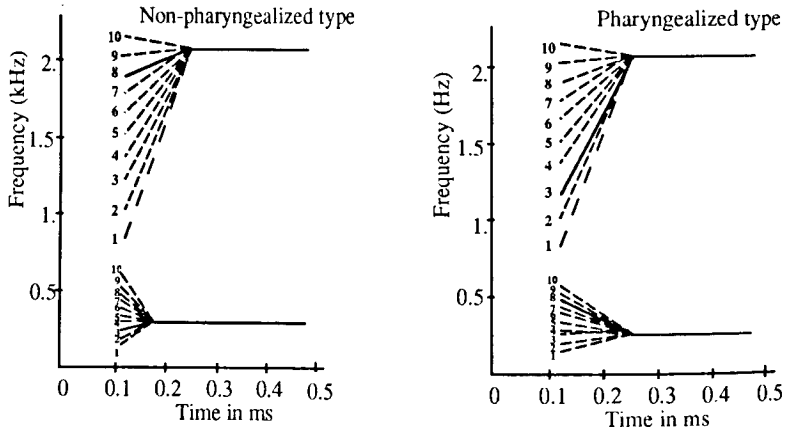


Figure 1. Schematic trajectories showing variations in $F1_0$ and $F2_0$ for the non-pharyngealized type continua (left) and pharyngealized-type continua (right). Reference stimuli are presented by solid lines.

All stimuli were of 470-ms duration and contained the same FO. Figure 1 provides a schematic representation of the continua: phayngealized and non-pharyngealized types.

2.1.2. Procedure and Subjects

The identification test consists of a randomization of the stimuli from the four series of continua. Each stimulus was presented 5 times. Intervals between stimuli were 2.5 s and between ten-stimulus blocks were 9 s.
The subjects were 11 graduate students (phonetics/linguistics). All subjects are Moroccan native speakers of Arabic and reported having normal hearing.

2.2. Results

Group identification functions for the F1 continua (both the pharyngealized type and the non-pharyngealized one) are presented in Figure 2. These functions show that variations in the onset F1 frequency are not effective in producing a perceived contrast between /sˤi:/ and /si:/: no crossing of boundaries occurs.

Figure 3 displays group identification functions for the F2 continuum (both pharyngealized and non-pharyngealized types). The functions show that the onset frequency of F2 is a critical acoustic property for the perception of the pharyngealized/non-pharyngealized contrast. Category boundaries were evaluated by interpolating the stimulus number at the 50 % crossover. The boundary is at 1276 Hz (near stimulus 4) for the pharyngealized-type continuum, and at 1773 Hz (near stimulus 8) for the non-pharyngealized-type continuum. The important difference between the two boundaries indicate that subjects would need an additional F2 onset lowering of 497 Hz to begin hearing [sˤi:] when F1 onset frequency is not appropriate for [sˤi:]. A t-test shows that such boundary difference is significant, $t(109)= 11.6, p < 0.0001$.

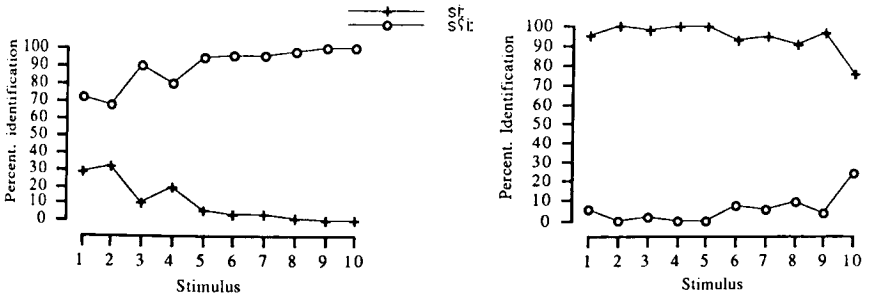


Figure 2. Identification functions for the F1 continua: pharyngealized type (left) and non-pharyngealized type (right).

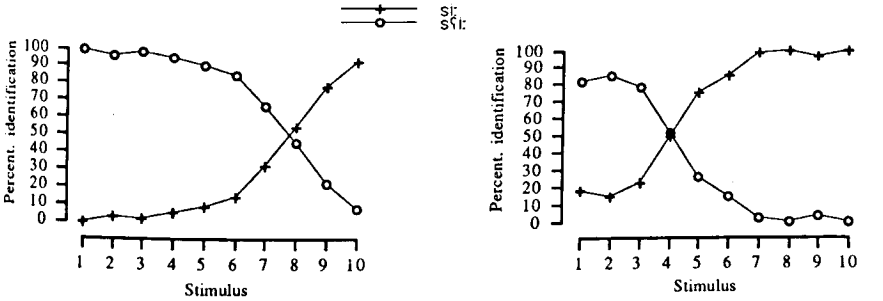


Figure3. Identification functions for the F2 continua: pharyngealized type (left) and non-pharyngealized type (right).

3. DISCRIMINATION TASK

3.1. Stimuli and procedure

Evidence for a trading relation between $F1_0$ and $F2_0$ was derived from the identification results of experiment 1. This would imply a perceptual equivalence between the two cues. The second experiment used a method testing the possibility of such perceptual equivalence [3]. This method investigates whether discimination performance would be differentially affected by the cooperation or conflict of the two cues along the phonetic dimension. Three comparing conditions of cue combination were used for this purpose: (1) one-cue condition, in which only $F2_0$ was varied; (2) cooperating two-cue condition, in which both $F1_0$ and $F2_0$ complemented each other phonetically (one member of each pair had one cue biased towards [si:], the other toward [s^{S}i:]); and (3) conflicting two-cue condition, in which the two cues cancelled each other. All ten stimuli of the continua were utilized, and each stimulus was paired with stimuli which were 4 steps apart from it on the spectral dimension (Δ 560 Hz). This amount approximates the amount of the boundary shift found in the identification experiment (497 Hz). The discrimination test (an AX task) was a randomised sequence of all possible stimulus comparisons repeated 5 times. The interval within each pair was 0.5 s and between succesive pairs 2.5 s.

3.2. Results

Group predicted discrimination scores derived from the identification data[1] are presented in Fig 4 and the corresponding obtained discrimination scores in Fig 5. A repeated-measure ANOVA with Conditions of comparison X Scores (predicted vs. obtained) was conducted on the results. There was no performance difference betwen predicted and observed scores across the cooperating cues condition, $F(1,131)=0.380$, $p=0.53$, and across the conflicting cues condition, $F(1,131)=0.345$, $p=0.55$. There was, however, a small significant effect for observed vs predicted scores across the one cue condition, $F(1,131)=5.390$, $p<0.02$. This difference indicates some ability to discriminate acoustic differences on a non-phonetic basis. This is also revealed by the fact that under the cooperating cues condition, the boundary-related peak was not well marked in the obtained scores compared with the predicted scores.

The above discrepancies are not damaging to the perceptual equivalence hypothesis, which is basically confirmed if discrimination performance in the cooperating cues condition is higher than performance in the conflicting cues condition. An ANOVA crossing Types of comparisons with Stimuls pairs was performed on the obtained data. Overall differences between types of comparisons were significant, $F(2, 180)= 188.93$, $p<0.0001$, and *post hoc* comparisons (*Fisher PLSD*) supported the perceptual equivalence prediction that the discriminability ordering would be: cooperating cues > one cue > conflicting cues.

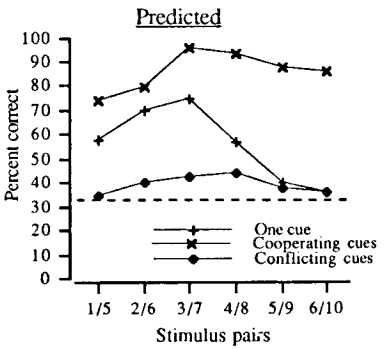


Figure 4. Group predicted discrimination scores in function of stimulus pairs and conditions of cue combination.

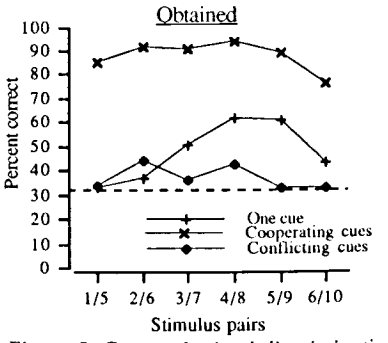


Figure 5. Group obtained discrimination scores in function of stimulus pairs and conditions of cue combination.

DISCUSSION

Most previous studies on trading relations have involved cues which are acoustically dissimilar and temporally separate and have generally supported the hypothesis that trading relations reflect phonetic perception which 'refers' to articulation/production [3, 4, 5, 6].

The present study replicates the findings of these studies for a new contrast involving two cues both spectral and temporally co-occuring. This would suggest the possibility of an interaction of some psychoacoustic origin, similar to that reported by [6,7], where the cues investigated consituted the same portion of the signal. The question that now arises: what is the phonetic and the auditory origins of the intergration of $F1_0$ and $F2_0$.

The phonetic explanation is based upon the hypothesis which explains trading relations with reference to articulation. Accordingly, the two cues are perceptually integrated because they are the result of the same articulatory gesture, i.e. pharyngealization. Such an articulatory rationale is not impossible; it is the most straighforward. In production of a pharyngealized consonant, the coarticulated /i:/ shows a high $F1_0$ (460 Hz) and a low $F2_0$ (1040 Hz). Both acoustic events are the result of a unitary articulation maneuver which includes: (1) a rearward movement of the back of the tongue towards the pharyngeal wall; and (2) a depression of the tongue's palatine dorsum [8]. The two movements result in a widened oral cavity and a reduced pharyngeal cavity. The high onset of F1 seems to be due to the reduction of the pharyngeal cavity, while the low onset of F2 to the widening of the oral cavity.

The psychoacoustic explanation is based on an auditory coherence account which proposes that listeners perceive the speech patterns of speech according to Gestalt principles. The principle that concerns us here is that of temporal proximity. $F1_0$ and $F2_0$ may cohere by virtue of their temporal proximity. They both bave onsets and offsets that are temporally simultaneous, i.e. they have the same duration (80-120 ms in the case of the vowel /i:/). Moreover they are very close in their frequencies with only a 580-Hz distance apart. This explanation is similar to that given in [9], where the harmonics of a vowel formant cohere by virtue of their temporal proximity.

[1] The following formula was used: $P_{corr}=[1+2(P_a-P_b)^2]/3$, where P_{corr} is the predicted probability of correct responses for a given stimulus pair, P_a is the obtained [s^{S}i:] responses to stimulus *a*, and P_b is the obtained [s^{S}i:] responses to stimulus *b* in the comparison. Chance level is at 0.33.

REFERENCES

[1] Yeou, M. (1995) "An investigation of locus equations as a source of information for consonantal place of articulation," *4th European Conference on Speech Comm. and Technology,* Madrid.

[2] Klatt, D.H. (1980) "A software for a cascade/parallel formant synthesizer," *JASA* 67: 971-995.

[3] Fitch, L., T. Halwes, & D.M. Erickson. (1980) "Perceptual equivalence of two acoustic cues for stop-consonant manner," *Perception and Psychophysics* 27 (4): 343-350.

[4] Best, C.T., B. Morrongiello, & R. Robinson (1981) "Perceptual equivalence of acoustic cues in speech and nonspeech perception," *Perception and Psychophysics* 29: 191-211.

[5] Repp, B.H. (1982) "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," *Psychological Bulletin* 92: 81-110.

[6] Repp, B.H. (1983) "Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization, " *Speech Communication* 2: 341-362.

[7] Polka, L. & Strange, W. (1985) "Perceptual equivalence of acoustic cues that differentiate /r/ and /l/," *JASA* 78: 1187-1197.

[8] Ghazeli, S. (1977) *Back Consonants and Backing Coarticulation in Arabic.* Ph.D. Dissertion, Texas University at Austin.

[9] Darwin, C.J. (1984) "Perceiving vowels in the presence of another sound: Constraints on formant perception," *JASA* 76: 1636-1647.

PHONATORY INSTABILITIES IN ALS AND MS: GRAPHIC AND QUANTITATIVE ANALYSES.

E. H. Buder[a], L. Hartelius[b], and E.A. Strand[a]
[a]University of Washington, USA
[b]University of Göteborg, Sweden

ABSTRACT

This paper reports on the use of a technique for examining long-term phonatory instabilities (flutter, tremor, and wow) in f_0 and dB of sustained phonations. The data are from subjects with amyotrophic lateral sclerosis (ALS) and multiple sclerosis (MS), and from gender- and age-matched controls. The poster displays the instabilities graphically. Initial results indicate excessive tremor in ALS and MS and excessive wow in ALS.

INTRODUCTION

A new technique [1] allows measurement of phonatory flutter, tremor and wow in f_0 and dB; Table 1 defines these domains for this study in terms of frequency range and minimum spectral magnitude. All of these phenomena are slower than cycle-to-cycle perturbations (jitter and shimmer), and are perceptible. The technique creates both graphic and statistical summaries.

Table 1. Domain Definitions

Domain	Frequency	Magnitude
Flutter	10 - 20 Hz	> 0.25
Tremor	2 - 10 Hz	> 1.00
Wow	0.2 - 2 Hz	> 1.50

This report applies the technique to 1) a group of four subjects with ALS (two men and two women) having mild to severe dysarthrias; 2) gender and age-matched subjects with MS having no discernible speech dysarthria; and 3) similarly matched control subjects. Table 2 provides the subjects' characteristics.

Previous work by our group has researched long-term phonatory instability of ALS ([2], [3], and [4]) and MS ([5], [6], and [7]) separately. The purpose of this study is to explore differences between small groups representing these populations, in order to guide hypothesis development for later work with larger datasets, and to demonstrate the technique.

Table 2. Subject Characteristics.

	Age	Yrs. Post-Diag	Dys-arthric
ALS, dysarthric			
Women	41	0.1	yes
	64	–	yes
Men	39	5.5	yes
	69	0.5	no
Age matched MS, non-dysarthric			
Women	40	1.0	no
	67	17.0	no
Men	40	6.0	no
	61	10.0	no
Age matched controls			
Women	39	–	no
	66	–	no
Men	40	–	no
	68	–	no

METHODS

Space does not permit a full description of the methods employed in this paper; details are reported elsewhere [1]. The technique includes the following steps: 1. a waveform of sustained phonation is digitized and analyzed for f_0

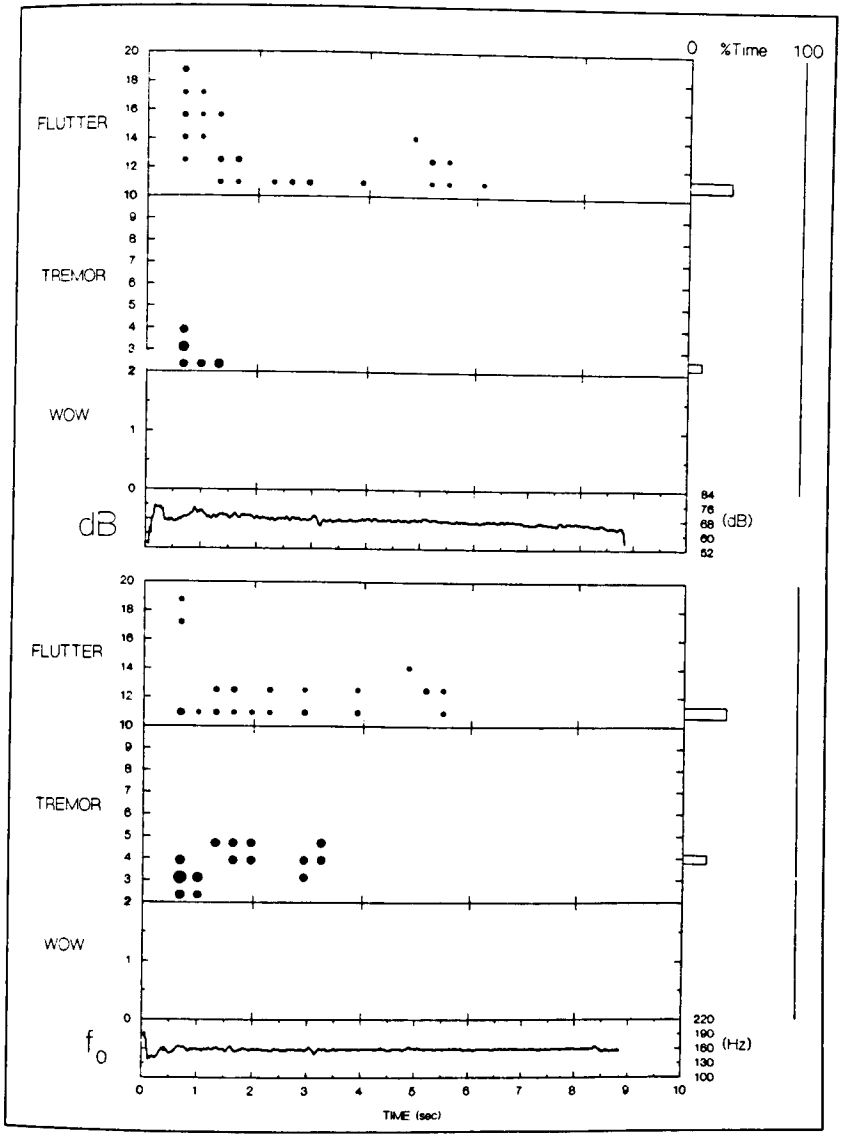


Figure 1. Graph of instabilities observed in a male subject with MS. Lower panels display f_0 and upper panels display dB. The display can be read like a spectrogram: time is on the abscissa, frequency on the ordinate, and magnitudes are displayed as dot size. Bars along the right edge indicate summary values of % time phonation as bar length and frequency of instability as vertical placement of the bar. See text for further details regarding analytical procedures.

and dB (performed in CSpeech [8] in our work), 2. the f_0 and dB data are smoothed and sampled at three different sampling rates — 200 Hz for flutter, 100 Hz for tremor, and 50 Hz for wow; 3. Fourier analyses are performed in successive frames using different transform sizes for the different domains — 0.64 s frames for flutter, 1.28 s frames for tremor, and 5.12 s frames for wow. The resulting magnitudes that pass the criteria listed in Table 1 are retained as observations for graphic or statistical analysis. We perform Step 3 and post-processing in Systat macros [9].

RESULTS

The observations can be graphed as in Figure 1, which provides a typical result from a (male) subject with MS: flutter in both f_0 and dB, some tremor in both parameters, but little or no wow. The instability observations as defined here can also be summarized by at least three different statistics: 1) the largest average magnitude at which a given instability was seen to occur, 2) the frequency at which this instability was observed, and 3) the percentage of total phonation time during which that instability was observed. The results for these variables are summarized within groups in Table 3. Gender is collapsed in this table, but the results are broken down by domain and parameter. For an initial exploration of effects associated with subject group, the data were also analyzed by a non-parametric Kruskal-Wallace analysis of variance. Figures 2 and 3 display some of the chief results obtained by this analysis, in which the data from f0 and dB parameters were pooled in order to maximize power. These are not the only significant results in the dataset, but isolate the effects that appear to be most strongly and uniquely associated with the different pathology groups. Figure 1 indicates that significant differences were found among all groups in percentage of time during which tremor was observed, and furthermore that the MS group is distinct from the controls in this measure. Figure 2 indicates that this pattern is slightly different in percent time of wow, showing that while there is again a significance of overall differences between groups, the MS group is in this case significantly lower than the ALS group. Together the results indicate that the distinction of domain of instability (e.g., tremor vs. wow) is helpful in isolating effects uniquely associated with the three conditions (ALS, MS, control).

Table 3. Phonatory instability measures (each group n =4).

		% Time	Freq.	Mag.
ALS, dysarthric				
flutter	f_0	77.6	10.9	0.59
	dB	27.2	11.3	0.39
tremor	f_0	53.6	2.5	1.59
	dB	16.6	2.3	1.39
wow	f_0	29.4	0.9	1.89
	dB	22.2	0.8	2.09
Age matched MS, non-dysarthric				
flutter	f_0	41.6	12.9	0.60
	dB	24.9	10.9	0.34
tremor	f_0	29.9	3.4	1.25
	dB	16.6	2.7	1.17
wow	f_0	3.6	0.4	1.79
	dB	3.6	0.6	1.51
Age matched controls				
flutter	f_0	24.1	11.7	0.35
	dB	3.9	11.5	0.32
tremor	f_0	12.0	3.1	1.23
	dB	5.0	3.9	1.51
wow	f_0	0.6	0.6	1.52
	dB	2.5	0.8	1.81

CONCLUSIONS

A technique has been presented for measuring, graphing, and summarizing phonatory instability in terms of two parameters (f_0 and dB) and three domains (flutter, tremor, and wow). The technique was applied to three groups (ALS, MS, and controls), and the feature allowing distinction of domain was found useful in discriminating samples from these populations. The phenomena identified by the technique are visually and perceptually clear, and may prove useful in clinical work. Research in the area is ongoing with larger populations ([1], [7]) allowing stronger statistical inference. Future research using the acoustic technique will focus on physiological and perceptual correlates.

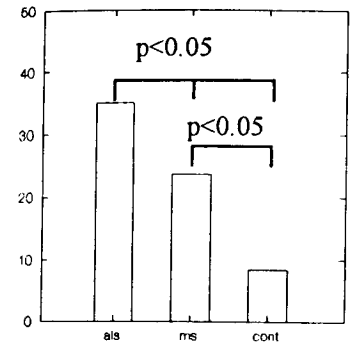


Figure 2. % Time Tremor in dB and f_0

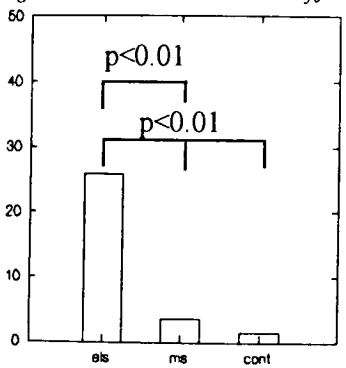


Figure 3. % Time Wow in dB and f_0

REFERENCES

[1] Buder, E.H., & Strand, E.A. (in submission). Acoustic, graphic, and statistical analyses of long-term phonatory instability in ALS.

[2] Buder, E.H., Iddings, S., & Strand, E.A. (1994, November). The development of phonatory tremor in ALS: A case study. Poster presented at the annual meeting of the American Speech-Language-Hearing Association, New Orleans, LA.

[3] Buder, E.H., Strand, E., & Iddings, S. (1994, March). A quantitative and graphic acoustic analysis of phonatory instability in ALS dysarthria. Poster presented at the Motor Speech Disorders Conference, Sedona, AZ.

[4] Buder, E.H., & Strand, E. A. (1993). Phonatory instability in ALS dysarthria: a case study. *The Journal of The Acoustical Society of America*, **94**, 1782 (abstract).

[5] Buder, E.H., & Hartelius, L. (1992, November). Quantifying long-term phonatory instability: Tremor due to Multiple Sclerosis. Paper presented at the annual meeting of the American Speech-Language-Hearing Association, San Antonio, TX.

[6] Hartelius, L., Nord, L., & Buder, E. H. (1995). Acoustic analysis of dysarthria associated with multiple sclerosis. *Clinical Linguistics and Phonetics*.

[7] Hartelius, L., Buder, E.H., & Strand, E.A. (in submission) Long term phonatory instability in individuals with Multiple Sclerosis.

[8] Milenkovic, P. (1994). CSpeech, Ver. 4.X. Author: Dept. of Electrical and Computer Engineering, University of Wisconsin, Madison.

[9] Wilkinson, L., (1990). *SYSTAT: the system for statistics*. Evanston, IL: SYSTAT, Inc.

SPEECH PRODUCTION CHARACTERISTICS OF CHILDREN FOLLOWING TRAUMATIC BRAIN INJURY

Thomas F. Campbell and Christine A. Dollaghan
University of Pittsburgh and The Children's Hospital of Pittsburgh, PA, USA

ABSTRACT

The purpose of the present investigation was to provide a comprehensive analysis of the speech production abilities of nine children with severe traumatic brain injury. Results revealed comparable phonemic-level skills in normal subjects and subjects who had suffered severe TBI approximately 13 months earlier. However, precision of articulation and suprasegmental aspects of speech production remained compromised in the majority of these TBI subjects.

INTRODUCTION

Despite considerable interest in the sequelae of pediatric brain injury, little information exists concerning these children's speech production skills. In previous investigations, few children have received comprehensive speech evaluations, and information on these children's speech production abilities, if mentioned at all, has been reported in general and anecdotal form. In addition, there are few published accounts of the changes in speech abilities during the recovery process.

The anecdotal reports that have appeared in recent years do suggest the presence of speech production deficits following TBI. One of the most commonly mentioned deficits is decreased speech intelligibility, generally presumed to result from motor planning and motor speech problems including dysarthria and apraxia of speech [1]. Prosodic and voice deficits also have been reported, with abnormalities in verbal fluency, speech rate, word and sentence stress, loudness, pitch and resonance among those mentioned [2].

Campbell & Dollaghan [3] examined the expressive language and speech skills of children and adolescents with TBI in a more comprehensive fashion than previous studies. Using developmentally appropriate measures of speech and language functioning in reasonably naturalistic tasks, language samples were obtained from these survivors seven times during a one-year period following the injury. The first sampling session occurred after the children and adolescents were discharged to a rehabilitation hospital and showed some evidence of intentional communication. The final session occurred 13 months later. Each brain-injured subject was age-matched with a normally developing non-injured child whose speech and language were sampled on the same schedule and with the same procedures. All of the children with TBI received less than 38 hours of speech and language treatment over the year following their injury.

The subjects in the Campbell & Dollaghan [3] investigation were nine children and adolescents ranging in age from 5:8 to 16:2 (years:months) at the time of injury. Four of the subjects were male and five were female. Each subject had English as a first language, and none had received speech, language, or psychological treatment prior to injury. In addition, all subjects had been functioning in normal classrooms prior to injury. Eight of the brain-injured subjects sustained closed head injuries from motor vehicle accidents, and the remaining child experienced an open head injury. All subjects were judged to be severely head-injured, meaning that they were unconscious for a minimum of 72 hours and received Glasgow Coma Scores of less than 11 (on a 15-point scale) for this time interval.

In Campbell and Dollaghan's initial report [3], data were presented on seven global measures of expressive speech and language output (including total number of utterances, total number of words, mean length of utterance in morphemes, percentage of complex utterances, percentage of utterances with mazes, and percentage of consonants correct). Not surprisingly, significant differences were found between the head-injured and non-injured groups on every one of the seven indices at the first sampling session, which occurred approximately one month post-injury. However, by the final sampling session, from 13-17 months post-injury, the groups differed on only one of these measures, with the brain-injured children producing significantly fewer utterances than their matched controls.

Results also showed changes in these subjects' speech production abilities over this 13-month period. There was a significant difference in the mean Percentage of Consonants Correct (PCC) [4] for the brain-injured (87%) and normal (98%) groups at the first sampling session. By the final session, there was no significant difference in the mean performance of TBI (95%) and normal (98%) groups. The PCCs of the individual normal subjects were quite stable across the sampling sessions, with performance generally above 90% correct. For the individual brain-injured subjects, there were measurable increases in PCC across the sampling sessions, with 7 of the 9 subjects producing at least 95% of their consonants correct at the final session. This result was somewhat surprising given the severity of these subjects' head injuries and suggests that their ability to correctly produce consonant phonemes in naturalistic conversation was reasonably close to that of their normal controls by 13-17 months following injury.

To further examine whether all brain-injured subjects actually reached the level of consonant articulation accuracy of their matched control subjects during the 13-month sampling period, a "normal performance quotient"[5] was calculated for each subject pair. The normal performance quotient was computed by dividing each brain-injured subject's PCC by the PCC of his or her matched control subject. A quotient of 1.0 would indicate that performance was equal to that of the uninjured subject. Results showed that six TBI subjects had performance quotients of 1.0 by the final session; all TBI subjects had performance quotients of at least .8 by sampling session 5, approximately three months post injury.

Based on these PCC results, it is tempting to conclude that the speech production deficits of most of these TBI children had resolved approximately one year after injury. However, PCC is a general index of consonant production and does not capture differences in articulatory precision or prosodic aspects of speech production. Clinical experience with these subjects suggested the need to examine articulation in more detail, as well as to consider other components of their speech production systems. Therefore, the purpose of the present investigation was to provide a more comprehensive analysis of the speech production abilities of these nine children with TBI and their age-matched normal control subjects approximately 1 year post injury.

METHODS

Subjects

The subjects for this investigation were the same 9 children with TBI and their age-matched normal control subjects described previously [3].

Speech Samples

A 12-minute conversational speech sample was obtained from each subject. For the subjects with TBI, the conversational samples were obtained from 13-17 months post injury. The conversational sample was the data set for the phonemic, phonetic, and

voice-prosody analyses. As described below, the subject's on-line narration of a 108-second video cartoon [6] was the corpus used for perceptual ratings of speech clarity and speaking rate.

Phonetic Transcription

The first 225 non-questionable words produced by each subject were transcribed phonetically into a microcomputer for analysis with the computer software programs entitled *Programs to Examine Phonetic and Phonological Evaluation Records (PEPPER)* [7]. Point-by-point agreement for phonemic transcription was above 90%.

Speech Analyses

To document the segmental and non-segmental characteristics of these subject's speech, a series of analyses were performed on the conversational samples. Segmental analyses included classification of phonemic and phonetic error types. Non-segmental analyses included ratings of prosodic and voice characteristics (e.g., phrasing, rate, word and sentence stress, loudness, pitch and vocal quality) using the Prosody-Voice Screening Profile [8] [9].

Finally, independent subjective ratings of speech clarity and speaking rate were obtained from naive listeners. This was accomplished by asking naive listeners to rate the 108-second video narration samples using direct magnitude estimation procedures [10]. Briefly, on two different occasions listeners judged a set of randomly ordered samples from brain-injured and control subjects with respect to speech clarity and speaking rate.

RESULTS

Results revealed comparable phonemic-level skills in normal subjects and subjects who had suffered TBI from 13-17 months earlier. As mentioned previously, the mean PCC value for the group with TBI was 93% while the mean PCC value for the normal group was 95%. Nearly all subjects with TBI produced more than 90% of consonants correctly, with only 1 subject's PCC falling below 85% correct. For the subjects with TBI, deletions of word-final consonants (typically fricatives and affricates) were the most common phonemic-level error; substitutions were noted in only two of the nine subjects. As indicated in Table 1, only one subject, whose PCC was 82%, was considered to have a phonemic-level deficit.

Phonetic-level errors were much more common and occurred in eight of the nine subjects with TBI. Inappropriate nasalization and lateralization of sibilants [s, z, ʃ] were common, and weak articulation and devoicing errors were also observed in approximately half of these subjects.

Table 1. Number and percentage of children with TBI displaying deficits in each component of speech production.

Speech Component	Number and % of Children Involved
Phonemic	1/9 (11%)
Phonetic	8/9 (89%)
Prosody-Voice	8/9 (89%)
Speech Clarity	5/9 (56%)
Speaking Rate	5/9 (56%)

Clinically significant deficits on the Prosody-Voice Screening Profile were found in all subjects but one at the final sampling session. Deficits in phrasing (word repetitions) were observed in seven of these subjects; six exhibited deficits in speaking rate and word/sentence stress. Four of the subjects displayed deficits in voice quality.

Finally, as described previously, naive listeners rated 108 second spontaneous samples obtained in a video narration condition with respect to speech clarity and speaking rate. Results revealed that the spontaneous speech of 5 of 9 BI subjects was rated significantly less clear than that of their normal control subjects; these same 5 BI subjects were rated as having significantly slower speaking rates than their control subjects.

CONCLUSION

The results of these analyses suggest that these TBI subjects experienced a significant recovery of phonemic-level skills over a period of approximately 13-17 months. However, precision of articulation and suprasegmental aspects of speech production remained compromised in the majority of these subjects such that naive listeners judged their communication skills to be significantly poorer than those of matched controls.

REFERENCES

[1] Thompson, C.K. (1988), Articulation Disorders in the Child with Neurogenic Pathology. In L.J. Lass, L.V. McReynolds, J.L. Northern, D.E. Yoder (Eds.): *Handbook of Speech and Language Pathology*, Philadelphia, PA., D.C. Becker, pp. 548-591.

[2] Ylvisaker, M. (1993). Communication outcome in children and adolescents with traumatic brain injury. *Journal of Neuropsychologic Rehabilitation*, vol. 3, pp. 367-387.

[3] Campbell, T.F. & Dollaghan, C.A. (1990), Expressive Language recovery in severely brain-injured children. *Journal of Speech and Hearing Disorders*, vol 55, pp. 567-581.

[4] Shriberg, L.D. & Kwiatkowski, J. (1982), Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders*, vol. 47, pp. 256-270.

[5] Bagnato, S.J. & Mayes, D.D. (1986), Patterns of developmental and behavioral progress for young brain-injured children during interdisciplinary intervention. *Developmental Neuropsychology*, vol. 2, pp. 213-240.

[6] Dollaghan, C.A., Campbell, T.F. & Tomlin, R. (1990), Video narration as a language sampling *context. Journal of Speech and Hearing Disorders*, vol. 55, pp. 582-590.

[7] Shriberg, L.D. (1986),)*User's Manual: Programs to Examine Phonetic and Phonologic Evaluation Records (PEPPER)*, Madison, WI., University of Wisconsin, Software Development and Distribution Center.

[8] Shriberg, L.D., Kwiatkowski, J. & Rasmussen, C. (1990), *Prosody-Voice Screening Profile [PVSP]: Scoring Forms and Training Materials*, Tucson, Arizona, Communication Skill Builders.

[9] Shriberg, L.D., Kwiatkowski, J., Rasmussen, C., Lof, G.L. & Miller, J.F. (1990), The Prosody-Voice Screening Profile (PVSP): Psychometric Data and Reference Information for children. *Technical Report No. 1.* Tucson, Arizona, Communication Skill Builders, pp. 1-54.

[10] Campbell, T.F. & Dollaghan, C.A. (1992), A method for obtaining listener judgments of spontaneously produced language: Social validation through direct magnitude estimation. *Topics in Language Disorders*, vol. 12, pp. 42-55.

THE STATUS OF SONORITY THEORY: EVIDENCE FROM SYLLABIFICATION IN APHASIC RECURRING UTTERANCES

*C. Code and M. J. Ball**
*University of Sydney, *University of Ulster*

ABSTRACT

The application of the sonority principle in syllabification is examined in non-lexical aphasic speech automatisms (recurring utterances). Syllabification was found to adhere to the sonority principle. These results are similar to those found with jargonaphasia. We discuss the location of syllabic sequencing principles in language organization, exploring the notion that sonority is either an artefact of the speech production process, or a hard-wired feature of phonological processing.

SONORITY

The term 'sonority' has had a long usage in phonology, and has in recent times been adopted by some syllable-based accounts of phonological theory. This concept has been used in three main ways: first in the description of sonority hierarchies (i.e. of segments), second in the description of sequencing within the syllable, and third in the ordering of segments within syllables.

Sonority has traditionally been defined from a perceptual viewpoint, in that the sonority of a sound is seen as its loudness relative to other sounds when length, stress and pitch are kept constant. Therefore, segments can be ordered along a sonority hierarchy, for example from least to most sonorant, stops, fricatives, affricates, nasals, liquids, glides, vowels.

The sonority sequencing principle (SSP) aims to account for segment ordering within syllables. This approach sees the syllable peak being highlighted by there being an increase of sonority from the syllable onset to the peak, and then a decrease from the peak to the coda. We expect, therefore, that in onsets, an initial obstruent would be followed by other segments increasing in sonority until we reach the peak (i.e. obstruent-nasal-liquid-glide-vowel, or O-N-L-G-V), while in syllable codas we expect the reverse ordering (V-G-L-N-O).

The syllable type with the greatest differentiation between onset and peak would be the OV type; the most favoured peak-coda type would be V-O, as here too there is the greatest sonority difference. The fact that some languages allow consonant clusters that do not follow the ordering set out above (e.g. OOV in "spy"), or allow syllables without onsets, etc., is accounted for by language specific phonotactic constraints.

SONORITY STUDIES IN APHASIA

The few previous studies of sonority and aphasia are reviewed in Christman [1]. The studies support the idea that there is an implicational hierarchy of syllable complexity, and that certain aspects of aphasic language breakdown may involve loss of control over more complex syllable structures.

Christman [1] notes that where the intended target is not clear, as with neologistic jargonaphasia, if neologisms still obey sonority constraints, "then we may find that they are not constructed with phonological abandon" (p225). She feels such results suggest that sonority is hard-wired in the brain in such a way that it survives extensive brain damage. Among other findings, her results showed that the overwhelming majority of both CV and VC patterns in the neologistic speech had an obstruent in the C position.

Christman comments that these results "support the notion of sonority as (1) a hard-wired component of the language system ...; (2) a mediator of phonological construction in all word forms, neologistic or otherwise ...; and (3) a useful metric in capturing the underlying phonological regularity of words that would otherwise appear to be somewhat randomly constructed ..." ([1] p.234).

RECURRING UTTERANCES

If we are to test the validity of the sonority approach further we need to explore other non-lexical forms in acquired neurological disorders. This study, therefore, examines aphasic non-lexical speech automatisms (recurring utterances).

While lexical speech automatisms are made up of recognizable words, and are syntactically correct structures in the majority of cases, non-lexical recurring utterances are mainly made up of reiterated and concatenated CV syllables (e.g. /bi bi/, /du du du/, /tu tu tu uuuu/). These utterances do not break the phonotactic constraints of the native language of the speaker.

The purpose of this study is to examine the syllable structure of non-lexical speech automatisms. Aphasics utilising speech automatisms tend to produce either one form only, or at the most very few different forms. For this study, therefore, we decided to make use of the data bases of British-English recurring utterances [2], and German recurring utterances [3]. The advantage of this approach is that we have immediate access to a relatively large amount of data from two different languages. As the majority of utterances recorded are of simple syllabic types, we are confident that the transcriptions are accurate enough for our analysis.

Method

The corpora for the study reported here were compiled from the studies noted above. This resulted in a total of 102 syllables for the English corpus, and 119 for the German corpus.

Our analysis requires the division of all syllables into demisyllables: that is the onset and peak of a syllable are assigned to an initial demisyllable, and the peak and coda (of the same syllable) are assigned to a final demisyllable. All demisyllables are further divided into utterance peripheral (initial or final), and embedded (initial or final), resulting in four categories: utterance initial (UI), embedded initial (EI), utterance final (UF), and embedded final (EF). Results of the analysis of the corpora are expressed in syllable shape (CV, CCV, V, VC), and demisyllable sonority profile (OV, NV, VO, etc.).

Results

The question we wished to address was the phonological make-up of demisyllables: both their phonological shape and their sonority profiles. Tables 1 and 2 show the demisyllable shapes for both the English and German recurring utterances, while Tables 3 and 4 show the sonority sequence patterns for the two groups.

Table 1. Syllable Patterns: English.

UI	EI	UF	EF
CV 20 69%	CV 60 82%	VC 4 14%	VC 5 7%
CCV 0	CCV 2 3%	V 25 86%	V 68 93%
V 9 31%	V 11 15%		
(n=29)	(n=73)	(n=29)	(n=73)

Table 2. Syllable Patterns: German.

UI	EI	UF	EF
CV 53 87%	CV 47 81%	VC 7 11%	VC 1 2%
CCV 1 2%	CCV 6 10%	V 54 89%	V 57 98%
V 7 11%	V 5 9%		
(n=61)	(n=58)	(n=61)	(n=58)

With the English recurring utterances, initial and embedded initial demisyllables were most frequently of the form CV. Of these CV types, OV was the most common demisyllable shape, with only NV scoring above 5% for both initial types, though GV occurred in a fair number of instances with embedded initial types. Only two initial consonant clusters occurred in the data, both of the OLV type. The V type occurred in 31% of the utterance initial, and 15% of the embedded initial demisyllables. Utterance final and embedded final demisyllables were overwhelmingly of the type V. The VO type was the only other variety found; no final clusters were found.

With the German recurring utterances, initial and embedded initial demisyllables were also most frequently of the form CV. For utterance initial this was followed by V and CCV, while for embedded initial the order was reversed. The most common type of CV was OV, followed by NV and LV. With clusters, the type found was OOV. These latter were phonetically /ts/, which could therefore be analysed alternatively as affricates (thus as further examples of OV). The choice of analysis does not

alter the overall balance of sonority profiles in any significant way. With final demisyllables, the V type is again overwhelmingly the favourite for both varieties. Small numbers of VC types occurred: split between VO and VN.

Table 3. Sonority Patterns: English.

UI	OV	17 S14, F13	59%
	NV	2	7%
	V	9	31%
	GV	1	3%
EI	OV	40 S36, F4	55%
	OLV	2	3%
	NV	8	11%
	V	11	15%
	GV	11	15%
	LV	1	1%
UF	VO	4 S0, F4	14%
	V	25	86%
EF	VO	5 S0, F5	7%
	V	68	93%

Tables 3 and 4 show the breakdown of the obstruent category into stops and fricatives, and demonstrates that the least sonorous obstruents, stops, are favoured in initial position, with fricatives favoured in final position (though the numbers here are small). This might be thought to agree with Clement's [4] view that final demisyllables show a minimal decrease in sonority.

Table 4. Sonority Patterns: German.

UI	OV	46 S35, F11	75%
	NV	7	11%
	V	7	11%
	OOV	1	2%
EI	OV	44 S39, F5	76%
	NV	2	3%
	V	5	9%
	LV	1	2%
	OOV	6	10%
UF	VO	3 S1, F2	5%
	VN	4	7%
	V	54	88%
EF	VO	1 S0, F1	2%
	V	57	98%

DISCUSSION

The essential findings of this study are that: i) the syllable shapes used in these non-lexical speech automatisms are generally of the simplest types phonotactically; ii)the sonority patterns of the demisyllables adhere closely to those predicted in sonority theory; and iii) no examples were found of language specific phonotactic ordering that supersede the Sonority Sequencing Principle.

We can examine the implications these results have for both sonority theory and the neural representation of recurring utterances. Non-lexical recurring utterances are not arbitrary but concatenated syllables governed by phonotactic constraints which adhere to the sonority structure of normal speech production and avoid language specific phonotactic possibilities that breach the sonority principle (e.g. /sC-/ for English, and /ʃC-/ for German). The non-lexical utterances appear to reflect articulatory simplification where only high frequency and motorically unmarked articulations taken from the phonetic inventory of the speaker's language are produced to conform to phonotactic rules. The fact that they do not break phonotactic constraints may suggest that they access a phonological output module the first time they are produced, and do not involve limbic-right hemisphere input. This conceptualization gains support from Sussman [5] who suggests that the reason phonotactic constraints are not violated in even the most severely aphasic patients, and syllabification unaffected by extensive brain damage, is because syllabification is 'hard-wired' in the left hemisphere of the brain.

However, we can consider other possibilities for the locus of syllabification control. For some left hemispherectomy patients reported in Code [6], specifically patients E.C. and N.F., the surgery was sufficiently radical to eliminate the possibility of the involvement of the remaining left subcortical structures in speech production. The phonotactic constraints of the language are not broken in the speech of these subjects, and syllabification is organized according to normal sonority. That is to say, removal of the left hemisphere in adulthood, while devastating for speech and language processing, does not appear to impede the syllabification of speech production.

This may suggest that Sussman's left hemisphere model for syllabification cannot be correct and that syllabification, if hard-wired, is hard-wired either subcortically or is diffusely represented throughout the brain. Some support for this comes indirectly from Ohala [7] who suggests that sonority is not an integral feature of phonological processing but merely an artefact of speech production. Christman [8] suggests that sonority may be 'well-distributed' both neurologically and linguistically, and may be accessed not simply at lexical levels to organize word syllabification or at sub-segmental levels to organize phoneme sequencing, but during different stages in speech production, including at the motor instantiation level.

Such a diffuse representation may reduce the strength of Sussman's argument that sonority is hard-wired in the left hemisphere and has a fully abstract cognitive representation, but implies that sonority enjoys no mental reality and is simply an inevitable by-product of speech production, an epiphenomenon of neurophysiology and the mechanico-inertial constraints of the speech production mechanism. This assumes, perhaps incorrectly, that 'hard-wired' is synonymous particularly with more focal representation. This may be why it survives even the most serious brain damage and even complete left hemispherectomy. However, surviving speech in global and in left hemispherectomy subjects is essentially non-propositional, formulaic, over-learnt and highly automatic, and such speech is probably not newly generated and phonologically processed each time it is produced [2]. The sonority and syllabification frame of such utterances may therefore be established during earlier, pre-lesion, propositional usage.

REFERENCES

[1] Christman, S. (1992b), "Uncovering phonological regularity in neologisms: contributions of sonority theory", *Clinical Linguistics & Phonetics*, vol. 6, pp. 219-47.

[2] Code, C. (1991), "Speech automatisms and recurring utterances", in Code, C. (ed.), *The Characteristics of Aphasia*, Hove: Lawrence Erlbaum. Pp. 155-78.

[3] Blanken, G., Wallesch, C-W. and Papagno, C. (1990), "Dissociations of language functions in aphasics with speech automatisms (recurring utterances)", *Cortex*, vol. 26, pp. 41-63.

[4] Clements, G. (1990), "The role of the sonority cycle in core syllabification", in Kingston, J. and Beckman, M. (eds), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, Cambridge: CUP.

[5] Sussman, H. (1984), "A neuronal model for syllable representation", *Brain and Language*, vol. 22, pp. 133-54.

[6] Code, C. (in press), "Speech from the isolated right hemisphere? Left hemispherectomy cases", in Code, C., Wallesch, C.-W., Joanette, Y. and Lecours, A.-R. (eds), *Classic Cases in Neuropsychology*, Hove: Lawrence Erlbaum.

[7] Ohala, J. (1990), "Alternatives to the sonority hierarchy for explaining segmental sequential constraints", paper presented at the *Parasession on the Syllable in Phonetics and Phonology*, Chicago Linguistics Society.

[8] Christman, S. (1992a) "Abstruse neologism formation: parallel processing revisited", *Clinical Linguistics & Phonetics*, vol. 6, pp. 65-76.

SPEAKING RATE AND LINGUISTIC PROCESSING SPEED IN CHILDREN AFTER ACQUIRED BRAIN INJURY

Christine A. Dollaghan and Thomas F. Campbell
University of Pittsburgh and Children's Hospital of Pittsburgh, Pittsburgh, PA, USA

ABSTRACT

Two studies were conducted to examine speaking rate following pediatric traumatic brain injury (TBI). In Study 1, five of nine subjects with severe TBI were found to have significantly slowed speaking rates, measured physically and perceptually, up to 13 months post injury. Study 2 revealed that reduced articulatory speed and increased pausing believed to be associated with linguistic processing difficulties may contribute independently to these speaking rate reductions.

INTRODUCTION

Interest in the speech and language abilities of children following traumatic brain injury (TBI) has grown significantly over the past decade. However, virtually no empirical evidence is available concerning one of the most commonly reported sequelae of TBI: significantly slowed speaking rate.

There are several plausible reasons why TBI might result in slowed speaking rates. First, complex motoric skills are known to be vulnerable to disruption by the diffuse damage characteristic of severe TBI [1]. Second, increased latency and slowed speed of response have been reported on a variety of neuro-psychological measures following TBI, particularly when processing demands are high [2]. This generally slowed processing speed could result in reduced speaking rate due to deficits in the processes needed to support such linguistic operations as lexical retrieval and syntactic formulation [3].

The present investigation consisted of two studies. In the first, we documented the magnitude of speech rate reductions longitudinally in nine children and adolescents with severe TBI. In Study 2, we examined the contributions of two potential influences on slowed speaking rate: reduced articulatory speed and increased pausing presumably reflective of deficits in linguistic processing.

STUDY 1: METHOD

Subjects

The subjects with TBI included 4 males and 5 females, ranging in age from 5;8 to 16;2 (years;months) at the time of injury. According to parent and school report, each subject had English as a first language, and none had received speech, language, or psychological treatment prior to injury. All subjects were classified as severely head-injured based on a post-injury period of at least 72 hours of unconsciousness, defined as a Glasgow Comas Score less than 11. Descriptions of these subjects' neurological, language, and cognitive profiles are available in Campbell and Dollaghan [4]. Each subject with TBI was matched with a normally developing control subject according to sex and chronological age at the time of injury (±3 months). By parent report, control subjects attended regular classrooms, and had no history of neurological disease or insult. Control subjects scored at or above their ages on a standardized vocabulary test.

Procedures

Speaking rate was measured in spontaneous speech samples obtained from each subject with TBI and his or her control subject during three different sampling sessions. The first sampling session occurred one month after the subject had been discharged from the acute care hospital and had begun attempting to communicate intentionally. The second and third sessions occurred seven and thirteen months after the first.

Speech samples were collected using an on-line video narration task [5] in which subjects describe the events occurring on a silent, 108-second videotaped cartoon. The video narration context represents a demanding language production task because of the time constraints imposed by the rapidly changing events on the cartoon. In addition, this task ensures the consistency (in rate, complexity, and sequence) of the events to be described across speakers and sampling sessions.

Utterances were recorded using a high quality audiotape recorder and external microphone for orthographic and phonetic transcription by trained research assistants.

Physical Measurement of Speaking Rate

Speaking rate was calculated for each subject using CSpeech, a computer assisted waveform analysis program [6]. Speaking rate, expressed in syllables per second, was calculated by dividing the total number of syllables produced (including interjections and other "maze" [7] phenomena) by the duration of the utterances, which included any silent pauses that occurred within utterance boundaries.

Perceptual Judgments of Speaking Rate

An important clinical question about any speaking rate reductions concerns their perceptual significance to naive listeners. To address this question, a direct magnitude estimation (DME) paradigm without modulus [8] was used to obtain perceptual ratings of speaking rate for individual subjects with TBI and control subjects at the final sampling session. To control for listener bias, the 18 video narration samples from the final session were dubbed from the original recordings onto a stimulus audiotape in random order for presentation to naive listeners. The direct magnitude estimates of speaking rate from each listener were converted to a common scale, and the means were computed for each subject with TBI and each control subject.

STUDY 1: RESULTS

The control group produced more syllables per second than the group with TBI at all three sessions, and average speaking rate changed little in either group over the three sampling sessions At Session 3, the control group's mean speaking rate was 4.74 syllables/s (SD = 1.07); the mean speaking rate in the group with TBI was 3.10 syllables/s (SD = 1.39).

Visual inspection of the speaking rate data within each subject pair, however, suggested marked differences between the TBI and control subjects in only five pairs at the final sampling session. Evaluating the significance of such visually determined differences is difficult. One approach is to calculate a "normal performance quotient" (NPQ) [9], defined as the ratio of the performance of each subject with TBI to that of his or her control subject. NPQs for the five subjects with TBI for whom speaking rate reductions were visually apparent were $\leq$ 0.6; none of the remaining four subjects with TBI had an NPQ lower than 0.78.

The naive listeners rated the final samples from five subjects with TBI as significantly slower than those of their matched controls. Importantly, these were the same five subjects whose slower physically measured speaking rates were visually apparent, and whose NPQs were 0.6 or lower.

The results of Study 1 confirmed slowed speech rate as a significant

sequela of TBI in some children and adolescents. Objective and subjective speaking rate measurements revealed significantly slowed speaking rates persisting more than one year post injury in five of these nine subjects with severe TBI.

STUDY 2

In Study 2, we examined the influence of two potential determinants of speaking rate. As noted earlier, generalized slowing of fine motor performance is a well-known outcome of TBI. Therefore, it is reasonable to speculate that slowed speaking rates might result from damage to the speech motor system. Alternatively, speaking rate could be slowed by reductions in the speed with which subjects conduct the cognitive-linguistic operations needed to access lexical items, construct syntactic frames and perform discourse processing operations. Perhaps most plausibly, speaking rate reductions in subjects with TBI could be associated with both sets of factors.

The influence of distinct motor-articulatory and cognitive-linguistic variables on speaking rate has been discussed by other investigators. One recent proposal [10] is that connected speech rate is determined by two factors: the speed of the articulators and the frequency and duration of silent pauses. These investigators suggested that "articulation rate" (i.e., number of syllables per second, calculated on runs of speech containing no pauses longer than 250 ms), may best reflect articulator speed and speech motoric performance. By contrast, they proposed that the frequency and duration of pauses longer than 250 ms may best reflect the operation of cognitive and linguistic factors.

In Study 2 we examined the interaction of these two factors in the speaking rates of our subjects. Specifically, we asked whether all subjects with slowed speaking rates exhibited slowed articulation rates in conjunction with increased percentages of within-utterance pause time, or whether these factors appeared to operate independently.

STUDY 2: METHOD

The first 50 syllables produced by each of the nine subjects with TBI at the final sampling session were analyzed. Duration and number of lexical syllables produced in runs of speech containing no pauses longer than 100 ms were calculated, yielding a measure of articulation rate in the form of average syllable duration. In addition, the duration of each pause longer than 100 ms was measured, yielding a measure of within-utterance pause time which was used to calculate the average percentage of time spent in silence within the utterance. Finally, to obtain a clinical judgment of speech-motor function, an experienced speech-language pathologist, blind to subject status, also independently rated samples from all subjects for the presence of dysarthria.

STUDY 2: RESULTS

Three of the five subjects with TBI who had slow speaking rates, based on the previous physical and perceptual measures, had average syllable durations more than two standard deviations above the average duration for the control group, clearly suggesting a contribution of speech motoric deficits to their slowed rates. Further confirmation of the existence of speech motor deficits in these three subjects was provided by the fact that these three subjects, and only these three, were rated as dysarthric by an independent speech-language pathologist. Four of the five subjects with TBI who were originally found to have slow speaking rates also had percentages of pause time more than two standard deviations above the average for the control group, suggesting that cognitive-linguistic factors contributed to their slow speaking rates.

Data from one subject illustrate the extent to which motoric and cognitive-linguistic contributions to slowed speaking rate can be dissociated. This subject's percentage of pause time was greater than that of the control subjects, but average syllable duration was not. It appears that this subject's slow connected speech rate can be attributed to linguistic formulation difficulties rather than to speech motor deficits, an interpretation that is bolstered by the fact that this subject was not rated as dysarthric.

The results of these exploratory analyses suggest that "articulation speed" and what might be called "cognitive-linguistic speed" may be dissociable in individual patients more than one year after TBI.

CONCLUSIONS

Study 1 provided the first empirical confirmation of the widespread clinical observation that slowed speech rate may be a significant sequela of TBI in children. Both objective and subjective speaking rate measurements revealed significantly slowed speaking rates persisting more than one year post injury in five of these nine severely brain-injured subjects. Study 2 revealed that these slowed speaking rates may not originate from a single source. Reductions in speaking rate after TBI may have different origins, and different implications.

REFERENCES

[1] Ewing-Cobbs, L.& Fletcher, J. (1990), "Neuropsychological assessment of traumatic brain injury in children." In E. Bigler (Ed.), *Traumatic brain injury*, Austin, Texas: Pro-Ed.

[2] Bawden, H., Knights, R. & Winogron, H. (1985), "Speeded performance following head injury in children", *Journal of Clinical and Experimental Neuropsychology*, vol. 7, pp. 39-54.

[3] Ewing-Cobbs, L., Levin, H., Eisenberg, H. & Fletcher, J. (1987), "Language functions following closed-head injury in children and adolescents", *Journal of Clinical and Experimental Neuropsychology*, vol. 9, pp. 575-592.

[4] Campbell, T. & Dollaghan, C. (1990), "Expressive language recovery in severely brain-injured children and adolescents", *Journal of Speech and Hearing Disorders*, vol. 55, pp. 567-581.

[5] Dollaghan, C., Campbell, T. & Tomlin, R. (1990), "Video narration as a language sampling context", *Journal of Speech and Hearing Disorders*, vol. 55, pp. 582-590.

[6] Milenkovic, P. (1988), *CSpeech* [computer program], Madison, WI: University of Wisconsin Electrical and Computer Engineering Department.

[7] Loban, W. (1976), *Language development*, Urbana, IL: National Council of Teachers of English.

[8] Schiavetti,N., Metz, D. & Sitler, R. (1981), "Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: evidence from a stuy of the hearing impaired", *Journal of Speech and Hearing Research*, vol. 24, pp. 441-445.

[9] Bagnato, S. & Mayes, D. (1986), "Patterns of developmental and behavioral progress for young brain-injured children during interdisciplinary intervention", *Developmental Neuropsychology*, vol. 2, pp. 213-240.

[10] Walker, J., Archibald, L., Cherniak, S. & Fish, B. (1992), "Articulation rate in 3- and 5-year-old children", *Journal of Speech and Hearing Research*, vol. 35, pp. 4-13.

AERODYNAMIC REGISTRATION OF SPEECH.

Gärda Ericsson and Árni Ingvarsson*
Dept. of Otorhinolaryngology, Linköping University Hospital, Sweden.
**Affiliated with Dept. of Linguistics, Stockholm University.*

ABSTRACT

Subtle deviations in articulatory behaviour can prohibit cleft palate patients from aquiring normal (non nasalized) speech after apparently successful surgery. This is not least a question of management of aerodynamics in speech. Registrations of nasal air flow and intraoral air pressure during therapy, with use of a dual oral-nasal Rothenberg mask, can be of great help for the patient as a pedagogical means in detecting and correcting faulty speech patterns.

INTRODUCTION

Patients who have established their speech production pattern under conditions incompatible with building up necessary intraoral air pressure have often adapted their articulatory strategies to these less favourable conditions. This is especially true of cleft palate patients who have needed secondary palatal surgery or who have had primary surgery late in life.

Not only deviant articulatory placement but also subtler deviations in the configuration of the vocal tract may prevent them from acquiring normal speech after apparently successful surgery. This is not the least a question of management of aerodynamics in speech. Warren (1986), [2] suggests that the regulation of aerodynamics is important in the development of speech production patterns. He further suggests that it governs the choice of compensatory behaviours and the programming of speech patterns in these patients. He proposes yet another approach for modifying these faulty speech patterns.

Subtle deviations in these patients have been described in detail and illustrated by spectrograms (Ericsson, 1987), [1]. Acoustic measurements have proved to be a valuable complement to usual clinical investigation of performance in these patients, and in evaluating the results of therapy. Understanding of all the information that is given in a spectrogram requires experience and possibility of consulting with experts in phonetics.

For the patient a spectrogram often might be difficult to interpret. Therefore it is of utmost importance to find a pedagogical means which the patient understands without difficulties. It is worth mentioning that these patients often don't actually hear their speech defects, even though they are aware that something must be wrong through comments from other people.

PRACTICAL PROCEDURES

We register with the aid of a dual oral-nasal Rothenberg mask, nasal air flow and intraoral air pressure during speech, before and after correction. The results are shown to the patient on the computer screen, which encourages and helps him to change his habit.

Thus, the patient can realise the results of articulatory changes with three senses: vision, hearing and feeling. To our knowledge, no one else is using simultaneous registrations of nasal air flow and intraoral air pressure as a pedagogical means during therapy.

EQUIPMENT

We use a dual oral-nasal Rothenberg mask. The openings in the mask is covered with fine wire mesh, to produce a pressure difference proportional to the air flow. However, the oral transducer is connected to a small tube which protrudes into the corner of the mouth.

Thus we measure intraoral pressure for labiodental sounds. The amplified output from the pressure transducers is connected to a 30 Hz lowpass filter to remove any speech frequencies from the registration. The filtered signal is digitised and recorded by SOUNDSWELL [3] , a speech recording and analysis software package. At the moment, we have no registration of speech, because it is most important to monitor the aerodynamics in the training situation for visual feedback. However, we are in the process of adding microphones to the Rothenberg masks for listening purposes and for extended evaluation of articulatory behaviour.

CASE DESCRIPTIONS.

Cleft palate patient

Registrations from a 16 year old boy are given in Figure 1a-c. He was born with bilateral cleft lip and alveolus and cleft palate. After lip operation the palate was operated before 2 years of age. Secondary operation with pharyngeal flap was performed when the boy was 10 years old, as the velopharyngeal function had proved to be insufficient.

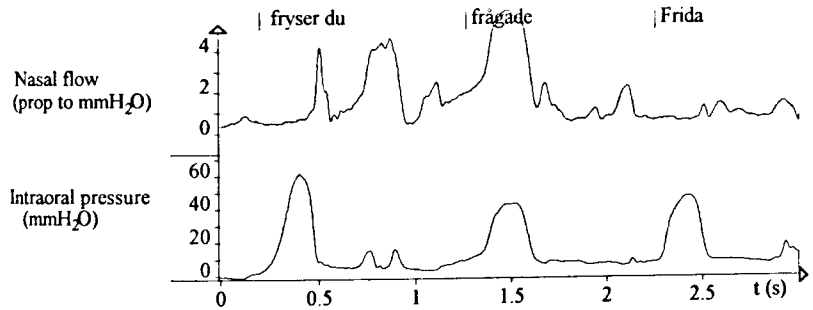


Figure 1 a shows nasal air flow and intraoral air pressure during pronunciation of the phrase: 'Fryser du frågade Frida?' (Are you cold? Frida asked.) before correction of speech.

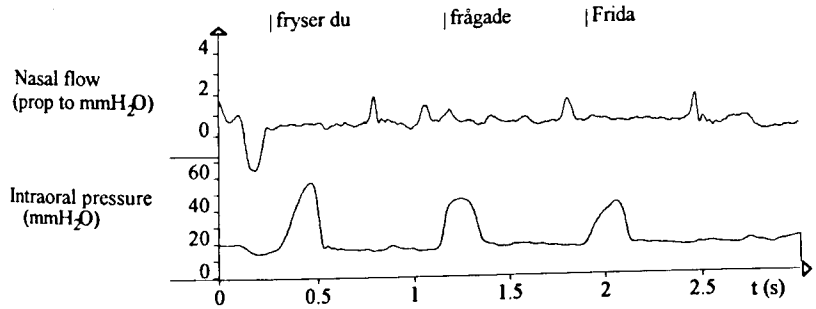


Figure 1 b shows the registrations of the same phrase immediately after correction.

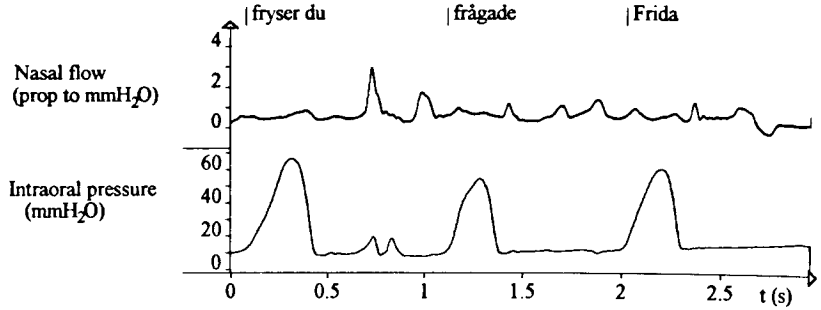


Figure 1 c shows the same phrase as spoken on first trial at a therapy session one month later.

Evident nasal escape can be noticed on Figure 1a. The relation in time of the nasal escape on /f/ to the intraoral air pressure built up on production of this sound can be studied. In the second and third registrations the nasal escape has diminished.With our equipment we can measure the air pressure in the oral cavity for sounds with articulatory seat in front of the little tube introduced in the corner of the mouth, thus bilabial and labiodental sounds. In training we have found it most convenient to start with these sounds, as the entire vocal tract is situated behind the place of articulation. Generalisation then often easily occurs to sounds of the same category on dental and velar places. Figure 2 shows the same phrase spoken by a normal speaking person.

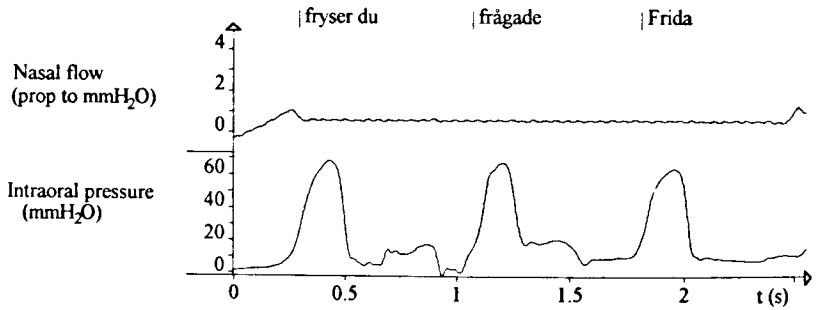


Figure 2 shows a registration from a normal speaker for comparison.

Patient with no history of cleft palate.

Figure 3 shows registrations from a 9 year old boy who was referred for nasal emission on /s/ in certain context. He had received speech therapy in school for some years. We could not find any nasal emission on /s/, but the patient seemed very aware of these sounds and prolonged them. On recording we noticed nasal emission sometimes but not always on /f/ and to lesser extent on /p/ and /b/. No history of cleft palate. The aerodynamic registrations show that the production of /f/ sometimes seemed to start with nasally directed airflow. On next visit, two weeks later, the nasal emission in the same utterance could not be noticed as shown in Figure 3b. The utterances shown here are bisyllables with final syllable stressed.

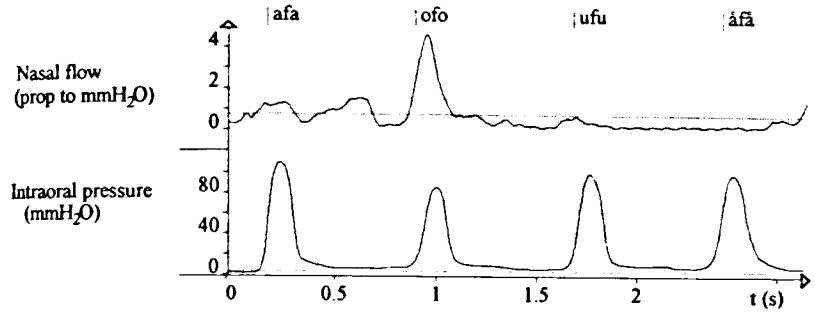


Figure 3a. Pretherapy registrations. Notice that nasal flow precedes the oral pressure build up for /f/ in "ofo".

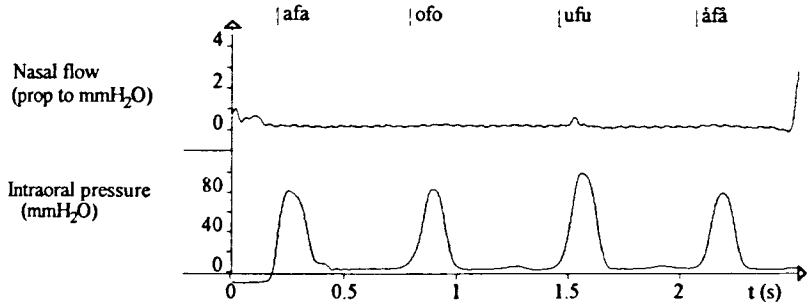


Figure 3 b. Posttherapy registration, two weeks later.

The boy and his mother found the aerodynamic registrations helpful in demonstrating the articulatory habit and the results of the changes during therapy. Later his tendency to prolong the /s/-sounds also disappeared, and his family members had spontaneously found his speech to be better.

REFERENCES

[1] Ericsson, G. (1987), *Analysis and Treatment of Cleft Palate Speech: Some Acoustic-Phonetic Observations*, Linköping University Medical Dissertations, No. 254.

[2] Warren, D. W. . (1986). Compensatory Speech Behaviour in Individuals with Cleft Palate: A Regulation/Control Phenomenon? *Cleft Palate J.*, vol. 23, pp. 251-260.

[3] *Soundswell user's manual*,1987-1994, Soundswell Music Acoustics HB

HEARING AID EVALUATION USING SPEECH PATTERN AUDIOMETRY

V. Hazan[1], G. Wilson[1], D. Howells[2], K. Reeve[1], D. Miller[2], E. Abberton[1,2], A. Fourcin[1,2]

[1] Dept of Phonetics and Linguistics, University College London, U.K.

[2] Laryngograph Ltd, London, UK

ABSTRACT

A new PC-based speech perception testing system, the Speech Pattern Audiometer (SPA) is described which aims to provide a simple and efficient clinical tool to assess listeners' ability to make use of acoustic cues to speech pattern contrasts for use in speech and language therapy clinics with those who are deaf or who have developmental or acquired speech perceptual disorders. This system constitutes a module of a complete speech and hearing workstation.

INTRODUCTION

In order to plan stage-appropriate speech and language therapy and to ensure the best possible hearing aid fitting, it is important to be able to reliably measure a deaf person's ability to make use of speech pattern information in the acoustic signal. In making such an evaluation, the contribution of other sources of information such as contextual information at the lexical, syntactic and semantic levels needs to be minimised, or at least carefully controlled. This makes sentence and word lists quite unsuitable for an evaluation of acoustic cue use.

DESCRIPTION OF THE SPA

In speech pattern audiometry, the ability to make use of acoustic cues to phonemic contrasts is assessed using a set of identification tests. Each test is based on a minimal pair composed of words of high-frequency of occurrence and easily represented by a picture. Each test assesses the ability to perceive a specific set of acoustic cues such as those which mark intonation contrasts, vowel quality, and voicing, place and manner of articulation in consonants. High-quality copy-syntheses based on tokens produced by a female British English speaker are prepared and the speech pattern cues under investigation are manipulated in a controlled way to construct a set of stimulus continua. Unlike tests based on natural speech, the speech pattern elements in the copy-synthesised words can be individually manipulated whilst maintaining a high degree of naturalness. By comparing performance in different test conditions in which the speech pattern cues are presented singly or in combination, it is possible to make a precise assessment of the speech pattern information used by a listener. As the tests assess the perception of common components of speech sounds, conclusions can be drawn about other speech sounds that are distinguished by the same speech pattern elements without explicitly having to test them.

Phonemic contrasts differ in terms of their speech pattern complexity and this is reflected in the age and stage of development at which they are acquired. By choosing a set of contrasts that spans these different levels of speech pattern complexity, it is possible to assess for example the stage of speech development reached by a child.

The stimuli from the continuum are presented to the listener in the form of a two-alternative forced-choice identification test: the listener hears a word and responds by touching the appropriate picture on a touch-sensitive response box. The ability to assign the sound to a specific phonemic category, which is assessed in this test, is central to the whole process of speech perception. The output of the test is a simple graph called an identification or

labelling function, which shows the percentage of responses of one label against the stimulus continuum. This graph can be economically described in terms of its gradient and phoneme boundary point, which are calculated using Maximum Likelihood Estimate (MLE) techniques [1]. Increasing confidence in identifying a contrast is marked by a sharpening of the identification function, and therefore by an increase in the measured gradient. These measures can be used in further statistical analyses, to look at evidence of change in performance over time or across different conditions.

The tests are *interactive* and the software which controls the test procedure determines the duration and complexity of each test on the basis of the client's ongoing performance. This is highly time- and cost-effective for the clinician, increases statistical reliability and ensures that clients are not frustrated by lengthy tests going beyond their ability. As each test takes only 3 to 4 minutes on average, it is possible to get a quantitative assessment of the perception of a range of contrasts within a twenty-minute session. Speech pattern audiometry has been evaluated in longitudinal studies of speech perception development with deaf children [2].

Test software and hardware

The SPA software has been implemented in Microsoft Windows. It includes facilities to store client records, to select and present minimal pair tests, to run interactive tests, to display and store test results in numerical and graphical form and obtain hard-copy printouts of the data. Facilities are also included to run basic statistics on the data.

The hardware required is a PC capable of running Microsoft Windows 3.1, fitted with a PCLX D/A card and a simple response box. In the absence of a response box, tests can be run using a mouse or joystick. As tests are usually carried out free-field aided, in a sound-treated room, a good quality amplifier and loudspeaker are also required. The sound level from the loudspeaker should be at the client's most comfortable listening level and must be monitored in each test session.

FIELD TRIALS

The sensitivity of both SPA and other audiometric tests for a particular evaluation of hearing aid performance was assessed in a study in collaboration with the Royal National Throat, Nose and Ear Hospital in London.

The test battery included the speech pattern tests described above and natural-speech based tests which take a similar analytic approach in assessing the use of acoustic information: the UCL 24-consonant VCV test [3] and the FAAF test [4].

Subjects

Subjects were four severely deaf listeners. All were regular hearing-aid users. They participated in the trial when they attended the clinic for their check-up approximately three months after being fitting with a new hearing aid.

Test battery

The listeners were each tested on a single day in two sound-proof rooms at the clinic. They were tested: (1) with their old aid; (2) with their new aid. In each condition, the following tests were presented in a "sound alone" condition (i.e. without lipreading).

Speech pattern tests

It was anticipated that for the clients selected, differences between hearing aids were most likely to be found in the perception of sounds cued by high-frequency patterns. Therefore 4 minimal pairs were selected which assessed the perception of place of articulation in initial plosives and fricatives. A initial-fricative voicing contrast was also included to assess the use of duration and low-frequency cues. The tests chosen were as follows (acoustic cues in parentheses).

- PEA-KEY (low-mid frequency burst and F2/F3 transitions)

- TEA-KEY (high-mid frequency burst and F2/F3 transitions)
- SUE-SHOE (mid-high frequency friction)
- SUE-ZOO (friction duration and presence/absence of voice bar)

24 consonant VCV test

The VCV test [3] investigates the perception of intervocalic consonants in nonsense words. An extended set of 24 consonants in a /a-a/ vowel environment was used. Each VCV was presented twice in random order. Listeners responded by writing a consonant on the answer sheet provided.

FAAF test

In the FAAF test [4] a test word is presented in a carrier sentence. The listener has to choose a response from four possible responses involving changes in the initial or final consonant contrast (e.g. "mail", "bail", "nail", "dale"). Each test contains 4 repetitions of 20 sets of stimuli. Results can be analysed to highlight the scores in voicing, place and manner of articulation.

Results

Speech pattern tests

Table 1: Identification function gradient for plosive contrasts.

	PEA - KEY		KEY - TEA	
Client	Old	New	Old	New
P001	-3.13	-1.87	*-1.31	0
P002	-5.00	-7.60	0	0.11
P003	-2.56	-3.51	*-0.28	0
P004	-1.07	-1.45	-0.39	-0.19

Table 2: Identification function gradient for fricative place and voicing contrasts.

	SHOE-SUE		SUE-ZOO	
Client	Old	New	Old	New
P001	-0.55	*-0.94	-1.12	-1.89
P002	-0.42	-0.44	-0.38	-0.24
P003	-0.93	-0.87	-3.51	-4.68
P004	-0.50	-0.34	-1.87	-1.56

The identification function gradients for each test with the two hearing aids are given above. The difference in gradient was judged as significant (marked by asterisk in Tables 1 and 2) if the gradients were a standard error apart.

For all listeners, steeper categorisation was obtained for contrasts marked by low-to-mid frequency cues (PEY-KEY and SUE-ZOO) than for mid-to-high frequency cued contrasts (SUE-SHOE and TEA-KEY).

Natural speech audiometry tests:

The difference in overall percent correct scores and in manner, voicing, and place of articulation scores obtained with the new versus old aid conditions for the two types of tests is presented in Tables 3 and 4.

Table 3: VCV test: Difference in scores between the new vs old aid.

	% total	% place	% voicing	% manner
P001	+18.7	+16.6	0	+10.0
P002	+38.8	+33.5	+33.5	+41.2
P003	- 4.2	+ 4.2	0	- 4.2
P004	0	- 4.2	+4.2	0.0

Table 4: FAAF test: difference in scores with the new vs old aid.

	% total	% place	% voicing	% manner
P001	+ 9	+ 4	+ 6	+ 3
P002	+ 3	+ 1	+ 3	- 5
P003	+ 5	+ 3	- 3	+ 3
P004	- 8	- 9	+ 3	- 3

Data analysis

P001 has six-frequency average (0.125 to 8 kHz) pure tone thresholds of 68 dBHL in the left ear and 38 dBHL in the right, with a flat configuration. She is labelling the PEA-KEY contrast sharply with both aids, but is showing better performance on the TEA-KEY contrast with the new aid, which suggests that this aid provides better frequency response at high frequencies. Sharper labelling of the fricative contrasts is also seen with the new aid. Finally, P001 shows better performance with the new aid in both natural speech tests.

P002 labels the PEA-KEY contrast, marked by low-to-mid frequency cues, very confidently with both aids, but shows poor performance with both aids on the TEA-KEY contrast and the two fricative contrasts. These results correlate with a lack of significant increase in performance with the new aid in the FAAF test. A significant increase in performance with the new aid is seen for the VCV test, but a careful examination of results suggests that this is due to particularly poor results with the old aid due to fatigue.

P003 has 6FA pure tone thresholds of 47 dBHL in the left ear and 50 dBHL in the right, with a very steep configuration (15dB loss at 1 kHz and 95 dB loss at 8 kHz). As might be expected, sharp categorisation was seen for the two low-to-mid frequency cued contrasts, PEA-KEY and SUE-ZOO. Poorer performance is seen for the mid-to-high frequency cued contrasts but significantly sharper labelling for the KEY-TEA contrast was obtained with the old aid. The natural speech tests do not conclusively show better performance with either aid.

P004 had 6FA pure tone thresholds of 109 dBHL in the left ear and 88 dBHL in the right ear. He too obtained steeper identification functions for the low-to-mid frequency cued tests. There was a trend to label sharply with the old aid for 3 of the tests. He also obtained higher scores on the FAAF test with the old aid.

DISCUSSION

A new clinical tool for speech perceptual assessment, the SPA, has been presented which is based on extensively tested techniques used in experimental phonetics research.

A clinical example is presented in which SPA was compared to other speech audiometry tests to assess the relative efficacy of two hearing aids for deaf clients. The VCV and FAAF tests evaluate the perception of a wide range of sounds. However, the VCV and FAAF feature-based performance measures (e.g. voicing and place correct) are still too general to provide much useful information for hearing aid fitting.

The speech pattern tests were successful in showing differences in performance which reflect the hearing aids' performance for low-frequency and high-frequency-based speech patterns. SPA had the further advantage of being quick and easy to administer, largely independent of vocabulary knowledge, and providing immediate scoring of results. This allows immediate feedback to be given to the client and results to be used within the session to make adjustments to hearing aid settings and try out new directions in rehabilitation. These tests therefore provide a valuable and powerful additional tool for audiological assessment.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contributions of Belinda Walker, Zoe Attard and Dr Barbara Cadge and colleagues at the RNTNE. The SPA development was funded by a DTI SMART award to Laryngograph Ltd. The clinical study was funded by a grant from the Nuffield Foundation (NUF-URB94).

REFERENCES

[1] Bock, K.D. & Jones, L.V. (1968), *The measurement and prediction of judgment and choice*, San Francisco: Holden Day.

[2] Hazan, V., Fourcin, A.J. & Abberton, E. (1991) Development of phonetic labeling in hearing-impaired children. *Ear and Hearing*, vol. 12, pp. 71-84.

[3] Rosen, S., Moore, B.C.J. & Fourcin, A.J. (1979) Lipreading with fundamental frequency information. *Proceedings of the Institute of Acoustics*, IA2, 5-8.

[4] Foster, J.R., & Haggard, M.P. (1979) (FAAF) An efficient analytical test of speech perception. *Proceedings of the Institute of Acoustics*, IA3, 9-12.

COORDINATION OF SPEECH PERCEPTION ABILITY OF COCHLEAR IMPLANT PATIENTS IN DIFFERENT LANGUAGES

Shizuo Hiki and Yumiko Fukuda***
**School of Human Sciences, Waseda University, Mikajima, Tokorozawa 359, Japan*
***Research Institute, National Rehabilitation Center for the Disabled, Namiki, Tokorozawa 359, Japan*

ABSTRACT

This paper reports a strategy for coordinating the results of evaluation of speech perception ability of cochlear implant patients for speech sounds in different languages. The strategy consists of the derivation of correction rules and the conversion of test results to allow coordination among different languages. Complementary effects in the combined use of multi-sensory channels were analyzed on IPA charts, and the data thus obtained were utilized in editing the computer program for conversion.

STRATEGY OF COORDINATION

Test results of speech perception ability of patients using cochlear implants or artificial inner ears in different languages were coordinated in two stages: the derivation of correction rules; and the conversion of test results.

Derivation of Correction Rules

In the first stage, by comparing the common parts of the phonetic and phonemic systems, and syllabic, word and sentence structures between different languages, correction rules A for objective language in regard to the reference language are derived. Also, by comparing the characteristic parts of the systems and structures of each language, correction rules B are derived *(Left of Figure 1)*.

Conversion of Test Results

In the second stage, test results in the objective language are converted to equivalent results in the reference language, using the correction rules derived in the first stage for the common parts and characteristic parts of both languages *(Right of Figure 1)*. Differences in the speech materials and types of test adopted in the evaluation experiments are also adjusted in the process of conversion.

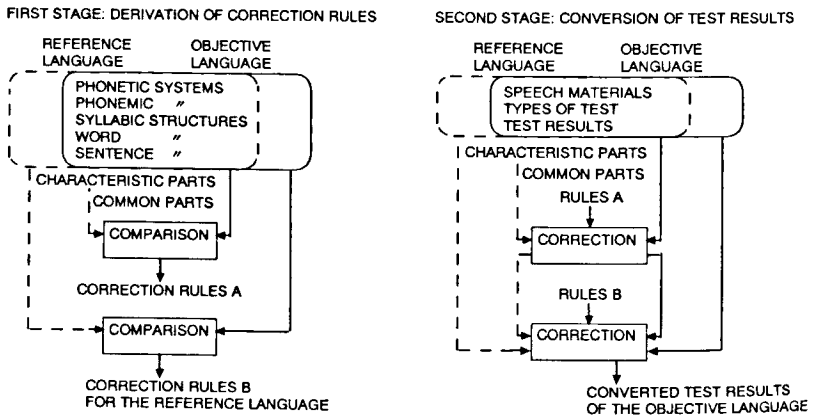


Figure 1.

MULTI-SENSORY PERCEPTION

An artificial inner ear is not sufficient in itself to enable for speech perception, but shows much better capability if combined with lip-reading. It is also necessary to evaluate the capability for speech perception through tactile sensation, as various types of tactile aids have been used in combination or substitutively with artificial inner ear or lip-reading.

Steps in Linguistic Processing

Information conveyed through each sensory channel can be analyzed commonly at the step of phonetic perception in the linguistic processing of speech. Outputs of optical analysis in visual organs and configurative analysis in tactile organs merge into the output of acoustical analysis in auditory organs at the step of the input of phonetic perception [1] *(Figure 2)*.

Analysis on the IPA Charts

The International Phonetic Alphabet (IPA) revised to 1989 was extended to include the classification of mouth shapes as well as speech sounds in the analysis of the transmission of speech information through auditory, visual or tactile sensory channels, so that the strategy is able to be applied to different languages in common.

Identifiable Groups of Phones

Auditory channel

For speech information transmitted auditorily, the amplitude-frequency ranges of the speech signal necessary for identifying each kind of consonant were derived from the data of an acoustical analysis and a perceptual test. Basic categories of consonants which are distinguishable from each other through the auditory channel were devised based on the consonant chart of the IPA *(Top of Table 1)*.

Visual channel

For visually transmitted speech information, identifiable groups of phones were predicted by taking into account the distinguishable places of articulation *(Middle of Table 1)*. The categorization was based on the stroboscopic observation of the mouth shape in the utterance [2]. Complementary effects between auditory and visual perceptions were estimated by referring to the data of hearing-impaired subjects using hearing aids or artificial inner ears [3].

Tactile channel

By using the same classification, groups of phones identifiable by tactile aids were predicted by referring to the speech perception data with a multi-channel vibro-tactile vocoder *(Bottom of Table 1)*.

EDITING COMPUTER PROGRAM

A computer program for converting results of evaluation experiments in different languages at various steps of linguistic processing is being developed by assembling the correction rules.

Reference and Objective Languages

In introducing the identifiable group of phones to the program, the speech sounds of Japanese is taken as the objective language, and that of English as the reference language, in the preliminary version.

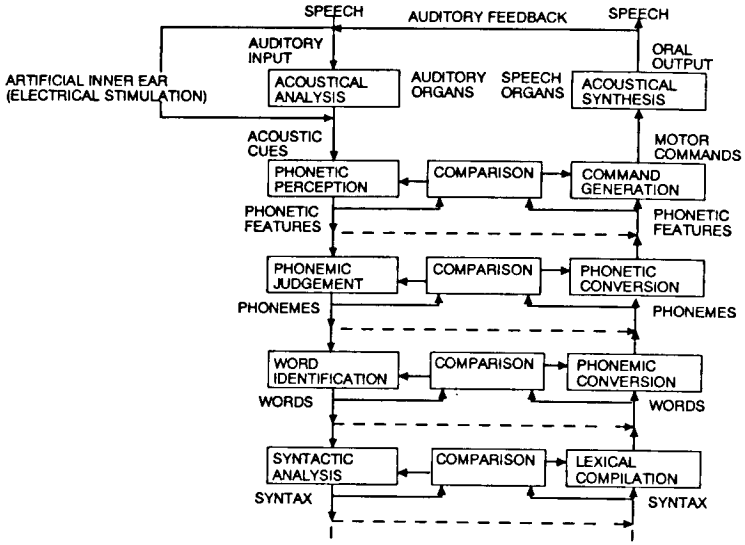
Type of Cochlear Implant Devices

The computer program for converting the test results also accommodates data obtained from various types of cochlear implant devices, in order to compare their capabilities when combined with lip-reading or tactile aids.

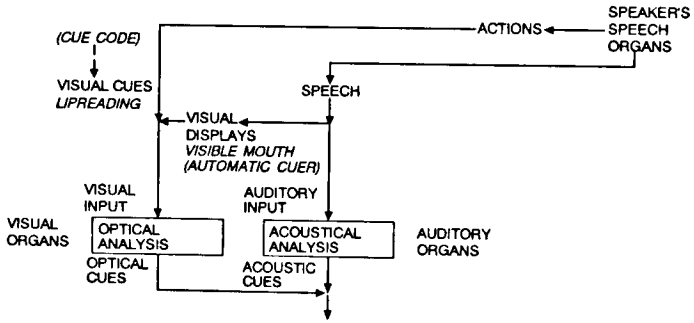
REFERENCES

[1] Hiki, Shizuo: Possibilities of compensating for defects in speech perception and production, Proceedings, 1994 International Conference on Spoken Language Processing, September, 1994, Yokohama, Japan, Vol. 4, pp. 2245-2252.
[2] Fukuda, Yumiko, and Hiki, Shizuo: Characteristics of the mouth shape in the production of Japanese: Stroboscopic observation, Journal of the Acoustical Society of Japan, pp. 75 - 91, 1982.
[3] Hiki, Shizuo, and Fukuda, Yumiko: Analysis of characteristics of speech perception by combined use of artificial hearing and visual/tactile sensation, Proceedings of 14th International Congress on Acoustics, Beijing, China, September, 1992, Vol. 3, H3-7.

STEPS IN SPEECH PERCEPTION AND PRODUCTION



VISUAL CUES IN SPEAKING ACTIONS AND STEPS FOR ACOUSTICAL / OPTICAL CONVERSION



STEPS FOR ACOUSTICAL / CONFIGURATIVE CONVERSION

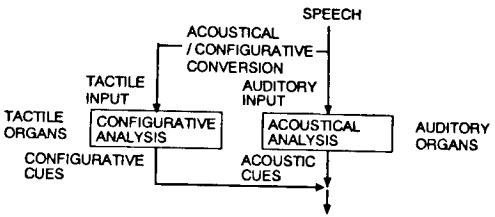
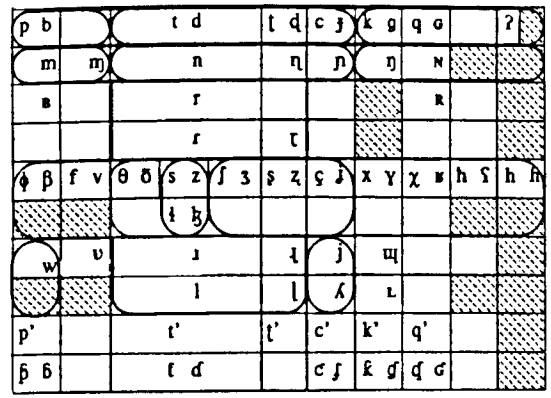


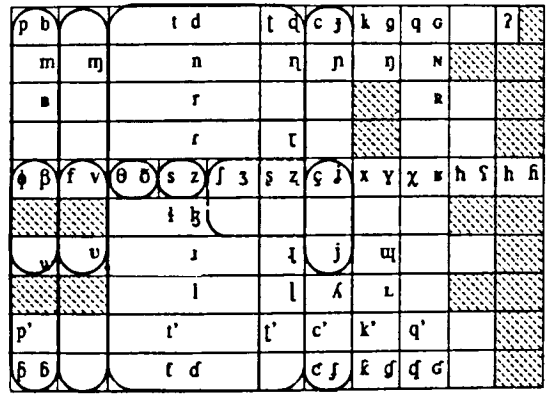
Figure 2.

Table 1.

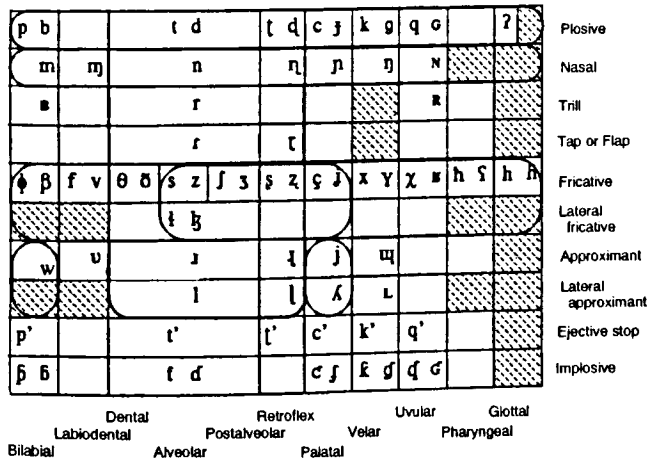
AUDITORY



VISUAL



TACTILE



A NEW DESCRIPTIVE SYSTEM FOR HAND SHAPES USED IN SIGNING BASED ON BIOMECHANICAL MODELING OF FINGER ACTIONS

Wako Ikehara, Emiko Kamikubo*, Shizuo Hiki** and Yumiko Fukuda****
**Graduate School of Human Sciences,*
***School of Human Sciences, Waseda University,*
Mikajima, Tokorozawa 359, Japan
****Research Institute, National Rehabilitation Center for the Disabled,*
Namiki, Tokorozawa 359, Japan

ABSTRACT

Based on the biomechanical modeling of finger actions, a new descriptive system for hand shapes used in signing is proposed. Hand shapes are classified systematically on a chart consisting of combinations of finger actions for changing the shapes and mutual relationships for each of the combinations of the active fingers. This chart serves as a language-independent framework for describing the hand shapes in manual communication.

PURPOSE

As part of the investigation of the coordination of descriptive systems proposed for different sign languages with biomechanical modeling of the hand and arm movements [1], classification of hand shapes is discussed based on the anatomical structure and control mechanism.

MODEL OF FINGER ACTIONS

Hand shapes are described based on the combination of finger actions for changing the shapes and mutual relationships, for each of the combinations of the active fingers.

The extension and flexion of the index, middle, ring and little fingers are modeled by the rotation of three kinds of joints (distal/proximal interphalangeal and metacarpophalangeal joints) caused by the contractions of seven kinds of muscles (flexor digiti minimi brevis, extensor digitorum/indicis/digiti minimi, flexor digitorum superficialis/profundus, and lumbrical muscles). While those of the thumb are modeled independently from the other fingers by the rotation of three kinds of joints (inter/metacarpophalangeal and carpometacarpal joints) caused by the contractions of four kinds of muscles (flexor/extensor pollicis longus/brevis muscles).

Abduction and adduction of the index, middle, ring and little fingers are modeled by the rotation of the metacarpophalangeal joint caused by the contraction of three kinds of muscles (dorsal interosseous, abductor digiti minimi and palmar interosseous muscles). Palmar abduction/adduction and radial abduction/ulnar adduction of the thumb are modeled by the rotation of the carpometacarpal joint caused by five kinds of muscles (abductor pollicis longus /brevis, opponens/adductor pollicis and flexor pollicis brevis muscles).

SYMBOL SYSTEM

By combining the above finger actions, hand shapes in signing can be described using the new system proposed here.

The descriptive system consists of symbols for the five fingers, marks denoting actions, and several marking conventions. The marks denote the actions of extended or flexed fingers for changing their shapes: extended, clenched, half-clenched, hooked angled, and half-angled, and those of extended fingers for changing their mutual relationship; and abducted, radial-abducted for the thumb, palmar-abducted and palmar-adducted also for the thumb, crossed and touched.

HAND SHAPE CHART

Theoretically possible hand shapes are classified in the form of a matrix consisting of the marks denoting finger actions for changing the shapes in columns, and for changing mutual relationships in rows, for each of the combinations of the active fingers. The hand shapes without and with the actions of the thumb are classified in chart A and B, respectively (Table1 and 2).

For simplicity of display, only the hand shape with the action indicated by the mark on the left-side of "/" is shown for the element in each column and row in these charts. The number of hand shapes involved in chart A and B is 215 and 572, respectively, and 789 in total.

DIFFICULTY IN ACTIONS

The number of hand shapes decreases to about 713 kinds, if some of the combinations are eliminated on account of being highly difficult to manipulate. (They are indicated by "−" in the charts.) The degree of the difficulty is estimated by analyzing the nature of control in the neuromuscular mechanisms of the hands, with regard to their original function of exerting forces to outside objects.

There are 76 kinds of shapes, which correspond to about 10% of total, in the criterion used here; and most of them are related to the actions of the middle finger, ring finger, index plus ring fingers and middle plus little fingers.

SHAPES USED IN SIGNING

These 713 kinds of hand shapes encompass most of the hand shapes found in the previously devised descriptive symbol systems for the manual alphabet and both traditional and simultaneous sign languages in America, 19 kinds of symbols proposed by Stokoe [2], 41 kinds of symbols by Friedman [3], 22 kinds of symbols by Liddel [4], 40 kinds of symbols by Sutton [5], 40 kinds of symbols in the SignFont [6]; in Sweden, 33 kinds of symbols by Bergman [7]; and in Japan, 54 kinds of symbols by Kanda [8] and 32 kinds of symbols by Yonekawa [9]. (They are indicated by "+" in the charts.)

The result is 78 different kinds of hand shapes, or about 10% of the total, but the remaining 80% of the hand shapes can be also utilized for finding a more effective system of the hand shapes in manual communication in general, as well as for describing the hand shapes in other sign languages.

APPLICATIONS

The descriptive symbol system of hand shapes proposed here is currently being applied in the computer simulation of signing gestures by combining it with models of arm and hand movements. It is also planned to apply it to the retrieval strings in electronic sign language dictionaries by modifying the system with regard to the degree of difficulty of recognition.

REFERENCES

[1] Fukuda, Y., Ikehara, W., Kamikubo, E., Hiki, S., An electronic dictionary of Japanese sign language: Design of system and organization of database, 1994 International Conference on Spoken Language Processing (ICSLP 94), September 18-22, 1994, Yokohama, Japan. Vol. 4, pp. 1987-1990.

[2] Stokoe, W.C., Sign Language Structure, Linstok Press, Inc., Silver Spring, Maryland, 1979.

[3] Friedman, L.A., Formational Properties of American Sign Language, On the Other Hand, Academic Press, New York, 1977.

[4] Liddell, S.K., Johnson, R.E., American Sign Language: The Phonological Base, Sign Language Studies 64, pp. 195-277, Linstok Press, Inc., Silver Spring, Maryland, 1989.

[5] Sutton, V.J., Sign Writing For Everyday Use, The Movement Shorthand Society, Inc., Boston, Massachusetts, 1981.

[6] SignFont Handbook, Edmark Corporation, Bellevue, Washington, 1989.

[7] Bergman, B., Studies in Swedish Sign Language, Institute of Linguistics University of Stockholm, 1982.

[8] Kanda, K. and Atari, H., Phonological notational system for Japanese Sign Language, Japanese Journal of Sign Linguistics, 12, pp. 31-39, 1991 (in Japanese).

[9] Yonekawa, A., A Study on Description of Sign Language, Meiji Shoin, Tokyo, 1984 (in Japanese).

Table 1. Hand shape chart A consisting of combinations of actions of the index, middle, ring and little fingers (215 shapes)

MARKS DENOTING ACTIONS FOR CHANGING SHAPES
(EXTENDED FINGERS) C: CLENCHED c: HALF-CLENCHED H: HOOKED A: ANGLED a: HALF-ANGLED
(FLEXED FINGERS) C': CLENCHED c': HALF-CLENCHED H': HOOKED A': ANGLED a': HALF-ANGLED

FOR CHANGING MUTUAL RELATIONSHIPS
(EXTENDED FINGERS) V: FINGER 2-5 ABDUCTED
FINGER 1 RADIAL-ABDUCTED
B: FINGER 1 PALMAR-ABDUCTED
d: FINGER 1 PALMAR-ADDUCTED
X: CROSS T: TOUCHED

SYMBOLS OF FINGERS
FINGER 2: INDEX, 3: MIDDLE, 4: RING, 5: LITTLE, 1: THUMB

MARKING CONVENTIONS
1. MARKS ATTACHED TO THE RIGHT OF SYMBOLS OF FINGERS
2. EXTENDED NOT MARKED
3. (OR): THE SAME ACTIONS FOR ALL FINGERS ENCIRCLED WITH A MARK
4. c' NOT MARKED FOR FINGER 1

+: USED IN THE SYMBOL SYSTEMS PROPOSED IN DIFFERENT SIGN LANGUAGES (78 SHAPES, ABOUT 10% OF THE TOTAL)
−: HIGHLY DIFFICULT TO MANIPULATE (76 SHAPES, ABOUT 10% OF THE TOTAL)

(FOR SIMPLICITY OF DISPLAY, ONLY THE HAND SHAPE WITH THE ACTION INDICATED BY THE MARK ON THE LEFT-SIDE OF / IS SHOWN FOR THE ELEMENT IN EACH COLUMN AND ROW.)

(ACTIVE FINGERS EXTENDED) MUTUAL RELATIONSHIP

	FINGER 2	V	X
FINGER 2	+2		
	c/H ++(2c		
	c'/H' 2(345c	2V(3V4V5H	
	A/a +2A		
SHAPE	A'/a' 2(345A		

	FINGER 0	V	X
FINGER 0	+(12345C		
	c/H		
	c'/H' +(2345c*	(2V3V4V5H***	
	A/a		
SHAPE	A'/a' (2345A**	(2V3V4V5a****	

	FINGER 3	V	X
FINGER 3	+23	+2V3 +2X3	+−3
	c/H +(23c	+(2V3H	(3c
	c'/H' 23(45c	2V3V(4V5H 2X3(45c	2c3(45c
	A/a ++(23A	(2V3a	3A
SHAPE	A'/a' 23(45A	2V3(45A 2X3(45A	−2A3(45A

	FINGER 4	V	X
FINGER 4	−24	−2V4	−34
	c/H (24c	−(2V4H	(34c
	c'/H' −23c45c	−2V3HV4V5H	2c345c
	A/a (24A	(2V4a	(34A
SHAPE	A'/a' −23A45A		−2A345A

	FINGER 4	V	X
FINGER 4	−3V4	+−4	
	c/H (3V4H	4c	
	c'/H' 2HV3V4V5H	(23c45c	(2V3HV4V5H
	A/a (3V4a	4A	
SHAPE	A'/a' −2A3V45A	−(23A45A	

	FINGER 5	V	X
FINGER 5	+25	+2V5	−35
	c/H (25c	(2V5H	−(35c
	c'/H' 2(34c5	2V(3V4HV5	−2c34c5
	A/a (25A	(2V5a	(35A
SHAPE	A'/a' 2(34A5		−2A34A5

	FINGER 5	V	X
FINGER 5	+5		
	c/H (5c		
	c'/H' (23c45	(2V3V4HV5	
	A/a (45A	(4V5a	
SHAPE	A'/a' (23A45	(23A4V5	

	FINGER 5	V	X
FINGER 5	+2345	+2V3V4V5	2X345
	c/H (2345c*	+(2V3V4V5H***	
	c'/H'		
	A/a (2345A**	(2V3V4V5a****	
SHAPE	A'/a'		

(ACTIVE FINGERS FLEXED) MUTUAL RELATIONSHIP

	FINGER 2	V	X
	345	3V4V5	
	c/H (345c	(3V4V5H	
	c'/H' 2c345	2HV3V4V5	
	A/a (345A	(3V4V5a	
SHAPE	A'/a' 2A345	2A3V4V5	

	FINGER 3	V	X
	245	−2V4V5	
	(245c	(2V4V5H	
	23c45	2V3H4V5	
	(245A	(2V4V5a	
	−23A45	−23A4V5	

	FINGER 4	V	X
	235	−2V3V5	−2X35
	(235c	(2V3V5H	
	234c5	2V3V4HV5	2X34c5
	(235A	(2V4V5a	
	−234A5	−2V34A5	−2X34A5

	FINGER 5	V	X
	+234	+2V3V4	2X34
	(234c	+(2V3V4H	
	2345c	2V3V4V5H	2X345c
	(234A	(2V3V4a	
	2345A	2V3V45A	2X345A

Table 2. Hand shape chart B consisting of actions of the thumb combined with chart A (572 shapes, 789 shapes in total of A and B)

(ACTIVE FINGERS EXTENDED) MUTUAL RELATIONSHIP

	FINGER 2	1V/B(d)	V	X
FINGER 2	12	+++1V(d)2		+1X2)
	c/H +(12c			
	c'/H' 12(345c	1V2(345c	1V2(3V4V5H	
	A/a 12A	++1V2A		
	A'/a' 12(345A	+1V2(345A		
	Tc/A 1T2)c			
SHAPE	Tc'/A' +1T345)c2			

	FINGER 1	1V/B(d)	V	X
FINGER 1	+1	+1V		
	c/H ++(1c			
	c'/H' 1(2345c	1V(2345c	1V(2V3V4V5H	
	A/a			
	A'/a' ++1(2345A	1V(2345A*****	+1B(2V3V4V5a******	
	Tc/A			
SHAPE	Tc'/A' 1T2345)c*******			

(ACTIVE FINGERS FLEXED) MUTUAL RELATIONSHIP

	FINGER 0	1V/B(d)	V	X
FINGER 0	+12345	+1V(d)2345	+1V(d)2V3V4V5	12X345
	c/H +(12345c		+(12V3V4V5H	
	c'/H'			
	A/a 1(2345A	+++1V(2345A*****	1B(2V3V4V5a******	
	A'/a'			
	Tc/A ++1T2345)c*******			
SHAPE	Tc'/A'			

	FINGER 3	1V/B(d)	V	X
FINGER 3	123	1V(d)23	+1V(d)2V3	12X3, +1X23)
	c/H (123c		+(12V3H	
	c'/H' 123(45c	1V23(45c	1V2V3V(4V5H	1V2X3(45c
	A/a 1(23A	+1V(23A	1B(2V3a	
	A'/a' 123(45A	1V23(45A	1V2V3(45A	1V2X3(45A
	Tc/A ++1T23)c			
SHAPE	Tc'/A' 1T45)c23		1T45)c2V3	1T45)c2X3

	FINGER 3	1V/B(d)	V	X
FINGER 3	−13	−1V(d)3		
	(13c			
	12c3(45	1V2c3(45	1V2H3(4V5	
	13A	1V3A		
	−12A3(45A	−1V2A3(45A		
	1T3)c			
	1T245)c3			

	FINGER 4	1V/B(d)	V	X
FINGER 4	−124	−1V(d)24	−1V(d)2V4	−134
	c/H (124c		−(12V4H	(134c
	c'/H' +123c45c	−1V23c45c	−1V2V3HV4V5H	12c34c5
	A/a 1(24A	1V(24A	1B(2V4a	1(34A
	A'/a' −123A45A	−1V23A45A		−12A34A5
	Tc/A 1T24)c			1T34)c
SHAPE	Tc'/A' −1T35)c24		1T35)c2V4	1T25)c34

	FINGER 4	1V/B(d)	V	X
	−1V(d)34	−1V(d)3V4	−14	−1V(d)4
		(13V4H	(14c	
	1V2c345c	1V2HV3V4V5H	1(23c45c	1V(23c45c 1V(2V3H4V5H
	1V(34A	1B(3V4a	14A	1V4A
	−12A34A5	−1V2A34A5	−1V2A3V45A	−1(23A45A −1V(23A45A
	1T34)c			1T4)c
	1T25)c34	1T25)c3V4		1T235)c4

	FINGER 5	1V/B(d)	V	X
FINGER 5	125	+1V(d)25	1V(d)2V5	−135
	c/H (125c		(12V5H	−(135c
	c'/H' 12(34c5	1V2(34c5	1V2V(3V4H5V	−12c34c5
	A/a 1(25A	1V(25A	1B(2V5a	1(35A
	A'/a' 12(34A5	1V2(34A5		−12A34A5
	Tc/A 1T25)c			−1T35)c
SHAPE	Tc'/A' +1T34)c25		1T34)c2V5	−1T24)c35

	FINGER 5	1V/B(d)	V	X
	−1V(d)35	−1V(d)3V5	145	1V(d)45 1V(d)4V5
		−(13V5H	(145c	(14V5H
	−1V2c34c5	−1V2HV3V4HV5	1(23c45	1V(23c45 1V(23H4V5
	−1V(35A	1B(3V5a	1(45A	1V(45A 1B(4V5a
	−1V2A34A5		1(23A45	1V(23A45 1V(23A4V5
	−1T35)c		1T4)c	
	1T24)c35	1T24)c3V5	+1T23)c45	1T23)c4V5

	FINGER 5	1V/B(d)	V	X
	15	+1V(d)5		
	(15c			
	1(23c45	1V(234c5	1V(2V3V4HV5	
	15A	1V5A		
	1(23A45	1V(234A5		
	1T5)c			
	+1T234)c5			

(ACTIVE FINGERS FLEXED) MUTUAL RELATIONSHIP

	FINGER 2	1V/B(d)	V	X
	1345	1V(d)345	1V(d)3V4V5	
	c/H (1345c	(13V4V5H		
	c'/H' 12c345	1V2c345	1V2HV3V4V5	
	A/a +1(345A	1V(345A	1B(3V4V5a	
	A'/a' 12A345	1V2A345	1V2A3V4V5	
	Tc/A 1T345)c			
SHAPE	Tc'/A' T1T2)c345		+1T2)c3V4V5	

	FINGER 3	1V/B(d)	V	X
	1245	1V(d)245	−1V(d)24V5	
	(1245c		(12V4V5H	
	123c45	1V23c45	+1V2V3HV4V5	
	1(245A	1V(245A	1B(2V4V5a	
	−123A45	−1V23A45	−1V23A4V5	
	1T245)c			
	+1T3)c245		1T3)c2V4V5	

	FINGER 4	1V/B(d)	V	X
	1235	1V(d)235+	−1V(d)2V3V5	−12X34
	(1235c		(12V3V5H	
	1234c5	1V234c5	+1V2V3V4HV5	−1V2X34c5
	1(235A	1V(235A	1B(2V3V5a	
	−1234A5	−1V234A5	+−1V2V34A5	1V2X34A5
	1T235)c			
	+1T4)c235		1T4)c2V3V5	

	FINGER 5	1V/B(d)	V	X
	1234	1V(d)234	+1V(d)2V3V4	12X34, +1X234)
	(1234c		(12V3V4H	
	12345c	1V2345c	1V2V3V4V5H	1V2X345c
	1(234A	1V(234A	1B(2V3V5a	
	12345A	1V2345A	1V2V3V4V5A	1V2X345A
	1r234)			
	+1T5)c234		1T5)c2V3V4	1T5)c2X34

SPEECH DEFECTS AFTER ORAL CANCER SURGERY
Functional and acoustic analyses of retrospective data

A-M. Korpijaakko-Huuhka[1], M. Lehtihalmes[2], A-L. Söderholm[3], A. Juvas[4], T. Jääskeläinen[4], C. Lindqvist[3].

[1]Dept. of Phonetics, University of Helsinki, [2]Dept. of Finnish, Saami and Logopedics, University of Oulu, [3]Dept. of Oral and Maxillofacial Surgery, Helsinki University Central Hospital, [4]Dept. of Oto-Rhino-Laryngology, Helsinki University Central Hospital, Finland

ABSTRACT

Fourteen speakers were studied for speech motor functions 10-144 months after resection of oral structures. Twelve speakers still had speech defects ranging from mild articulatory problems to severe unintelligibility. The most severe defects correlated with the most deviant F1/F2 distributions showing either generally centralised formant values or lowered F2-frequencies, especially for the front vowels.

INTRODUCTION

Surgical operations in the oral structures alter speech and other oral motor functions. However, our knowledge of long-lasting effects of oral cancer surgery is limited [1,2]. Also, the need for systematic and multidiciplinary evaluation of patients with operated oral cancer has not been recognised until recently in Finland. The present study belongs to the retrospective part of a research project which aims to evaluate functional consequences of oral cancer treatment.

The aim of the present study was to examine the correlation between intelligibility ratings and acoustic features of vowels [1,3,4] To reach this aim clinically evaluated speech defects were compared with psychoacoustical F1/ F2 charts.

METHODS

Fourteen speakers (8 women and 6 men) were studied about four years (range 10-144 months) after resection of oral structures due to oral cancer. The mean age of the speakers was 59 years (range 22-82). The lesions involved the mandible in seven speakers, the maxilla in three, the tongue in two, and the floor of mouth in two speakers. Lesions after surgery were largest following resections of the bony structures in mandible and maxilla. Nine subjects had radiation therapy postoperatively and also nine subjects had mandibular reconstructions or obturator prosthesis to compensate the structural defects.

Clinical evaluation of speech and oral motor functions

Two qualified speech therapists estimated the adequacy of speakers' oral motor functions and speech intelligibility with the Frenchay Dysarthria Assessment (FDA) [5], a sentence repetition task from the Finnish version of Speech Examination [6,7], a text reading task [8], a nasality assessment [9], and finally, with a story telling task [10]. Oral motor examinations were videotaped for reliability testing, and speech samples were audiotaped in a soundproof room with a Revox-A77 tape recorder.

The data consists of the results from FDA profiles converted to points (max. 224 points). In addition, the severity of speech defect was estimated on a 4-point scale (0=none and 3=severe) by the first two authors.

Acoustic analysis of vowel quality

A task of reading the eight Finnish vowels (both short and long) was also recorded. The test items consisted of 96 words in which each vowel type occurred in six different contexts (e.g. *tippa* 'drop', *piha* 'courtyard'; *piina* 'agony', *tiili* 'brick'). The speakers were asked to read each word three times without interword pauses, i.e. as a sentence-like utterance (e.g. *tippatippatippa*) [11].

To estimate deteriorations in vowel quality, a psychoacoustical F1/ F2 chart was applied utilising the LPC-analysis in the ISA-program [12, 13]. The formants were measured at the temporal midpoint (when possible) of the first vowel of the second word in each three-word sequence. The mean values of the six measurements per each vowel class were then calculated in order to compute the final F1/F2 charts for each speaker.

RESULTS

All speakers claimed that, acutely after the operation, they had had speech difficulties. At the time of the present study all but two speakers still had articulation deficits and/or hypernasal voice quality. In four speakers, a problem severe enough (ratings 2 or 3) to affect speech intelligibility was detected.

Clinical findings

The most severe oral motor and speech defects were found in speakers with partial glossectomy or mandible resections due to the fact that, in many cases, the operation reduced mobility of lingual muscles in addition to the structural changes in the oral cavity. Nasal voice quality characterised speakers with maxilla resections due to insufficient velopharyngeal functions, even when they were wearing an obturator prosthesis. Table 1 shows the essential findings of the functional analysis.

Table 1. Mean scores from FDA (max. scores are mentioned in parentheses). Mand, Max, Tong and Floor refer to lesions of mandible, maxilla, tongue, and floor of mouth, respectively. Palate, Intell and Total refer to palatal functions, intelligibility scores, and total scores of the functional assessment, respectively.

	Palate (24)	Tong (48)	Intell (24)	Total (224)
Mand	22.4	27.6	20.4	183.7
Max	8.0	45.0	20.7	192.7
Tong	23.0	28.0	18.0	188.0
Floor	23.0	32.5	21.0	197.0

F1/F2 patterns

Different patterns of variability in formant frequencies were observed when speakers' F1/F2 charts were compared with a normative pattern (see figure 1) [14]. First, the relative distances between vowel classes were found to correspond the normal pattern in five cases. In four speakers the speech defect was rated either non-existent or mild. The fifth speaker was the oldest in this sample (82 years). She also produced a normal vowel pattern, but her moderate speech defect (rating 2) resulted both from effortful articulation due to lateral fixation of the tongue tip, and from phonatory problems; the latter was most obviously related to her high age

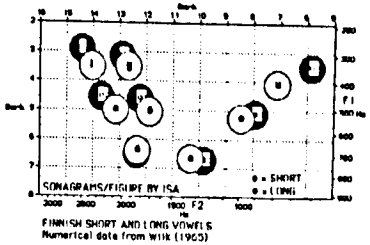


Figure 1. The psychoacoustical space of normal Finnish vowels [14, 15].

Centralisation of formants was found clearly in seven speakers, who also had more severe speech defects, in the average. However, a general F1/F2 reduction of formant variability was observed in two speakers only. The one had a moderate speech defect and the other a severe one (rating 3, see figure 2) as a result from hypernasality and severe reduction of tongue movements, respectively.

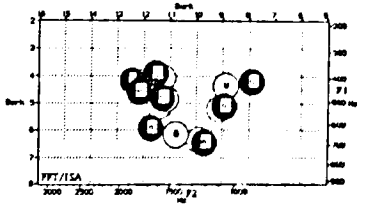


Figure 2. General centralisation of F1/F2 frequencies due to reduction of tongue movements after mandibular resection (case 14).

A third F1/F2 pattern was related to lowered F2-values for front vowels in two cases. Both speakers had difficulties in moving the tongue towards lips in clinical examination, but the one (with mild speech defect) was able to raise the tongue while the other (with severe speech defect) was not (case 5, figure 3).

Figure 3. Lowered F2 frequencies due to hemiglossectomy (case 5).

A fourth pattern was observed in three cases. It was characterised by reduced variability of F1-values (see figure 4), resulting from the speakers' difficulties in raising the tongue. This reduced the accuracy of pronunciation of tip-alveolar phonemes, especially the Finnish liquid [r]. On the whole, speech defect was rated mild in these three speakers.

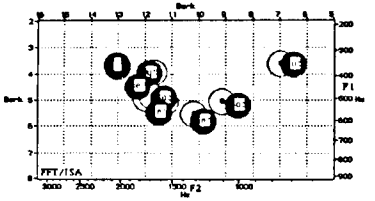


Figure 4. Reduced F1 frequencies after resection of the floor of the mouth (case 13).

In two cases, severe nasality decreased the reliability of the measurement of formants. Therefore these results are not presented here [1]. The speech defect was rated mild in both of them.

DISCUSSION

Long-lasting speech deficits and other oral motor disorders [16] were found in twelve of the fourteen subjects operated on oral cancer. In eight speakers, the mild speech defect had only minor effects on speech intelligibility. In four subjects a more severe deficit was found.

Speech intelligibility can be affected by several factors, e.g. phonatory problems, consonant distortions, hypernasality and deviations in vowel quality. In the present study, only vowel quality was examined. In spite of evident individual variation in acoustic data, four different patterns of F1/F2 charts were found. In one group, formant frequencies matched closely the normative chart. The speech difficulties were also mild in those speakers, expect for one subject with age-related phonatory changes. The restricted vertical tongue movement reduced the variability of F1 but did not affect speech intelligibility. However, consonant pronunciation was less accurate in that group. The most deviant F1/F2 distributions were marked lowering of F2 frequencies for front vowels or a general reduction of the relative distances between vowel classes on the F1/F2 chart. These formant patterns were found in four speakers with moderate to severe speech defects due to restricted mobility of the tongue. The relationship between speech intelligibility and F2 values has been reported also in dysarthric speakers [17]. In addition to deviant vowel pattern, the speakers in our study presented various problems in consonant pronunciation as well as hypernasality.

The number of speakers contributing each group was small in this preliminary study. However, the results support some earlier findings that front vowels are most vulnerable in oral surgery [1,4]. On the other hand, the speakers seem to sustain the formant pattern for back vowels better than for the front vowels, even with minor movements of the posterior part of the tongue.

To improve the external validity of our results, a larger group of subjects will be studied both retrospectively and prospectively during the present research.

REFERENCES

[1] Leonard, R., Goodrich, S.,, McMenamin, P. & Donald, P. (1992), Differentiation of speakers with glossectomies by acoustic and perceptual measures. *American Journal of Speech-Language Pathology* 1, (4), 56-63.

[2] Pauloski, B.R., Logemann, J.A., Rademaker, A.W., McConnel, F.M.S., Heiser, M.A., Cardinale, S., et al. (1993), Speech and swallowing function after anterior tongue and floor of mouth resection with distal flap reconstruction. *Journal of Speech and Hearing Research* 36, 267-276.

[3] Tobey, E.A. & Lincks, J. (1989), Acoustic analyses of speech changes after maxillectomy and prosthodontic management. *Journal of Prosthetic Dentistry* 62, 449-455.

[4] Leonard, R. & Gillis, R. (1982), Effects of a prosthetic tongue on vowel intelligibility and food management in a patient with total glossectomy. *Journal of Speech and Hearing Disorders* 47, 25-30.

[5] Enderby, P. (1981) *Frenchay Dysarthria Assessment.* San Diego: College-Hill Press.

[6] Keller, E. (1990), *Instructions for scoring the Speech Examination (SE).* Version 2.0, August. (Unpublished manuscript).

[7] Werner, S., Tuomainen, J. & Lehtihalmes, M. (1990), *The Speech Examination (SE).* (Unpublished Finnish translation).

[8] *The Principles of the International Phonetic Association* (1965), University College, London.

[9] Haapanen, M-L. (1992), Factors affecting speech in patients with isolated cleft palate. *Scandinavian Journal of Plastic and Reconstructive Surgery*, Supplementum 26.

[10] Korpijaakko-Huuhka, A-M. & Aulanko, R. (1994), Auditory and acoustic analyses of prosody in clinical evaluation of narrative speech. *Proceedings of the Third Congress of the International Clinical Phonetics and Linguistics Association, 9—11 August 1993, Helsinki*, (eds. R. Aulanko & A-M. Korpijaakko-Huuhka). Publications of the Department of Phonetics, University of Helsinki 39, 91-98.

[11] Iivonen, A. & Laukkanen, A.-M. (1993), Explanations for the qualitative variation of Finnish vowels. *Studies in Logopedics and Phonetics 4*, (eds. A. Iivonen & M. Lehtihalmes). Publications of the Department of Phonetics, University of Helsinki, Series B: Phonetics, Logopedics and Speech Communication 5, 29-54.

[12] Iivonen, A. & Toivonen, R. (1989) Simulation of the psycho-acoustical vowel space for linguistic applications. *Eurospeech 89. European Conference on Speech Communication and Technology*, Paris, September 1989, (eds. J.P. Tubach & J.J. Mariani),, 289-292.

[13] Iivonen, A. (1992), Articulatory vowel gesture presented in a psychoacoustical F1/F2-space. *Studies in Logopedics and Phonetics 3* .(eds. R. Aulanko & M. Lehtihalmes). Publications of the Department of Phonetics, University of Helsinki, Series B: Phonetics, Logopedics and Speech Communication 4, 19-45.

[14] Iivonen, A. (1990), An outline of an acoustical vowel data base. *Studies in Logopedics and Phonetics 1.* (eds. M. Leiwo & R. Aulanko). Publications of the Department of Phonetics, University of Helsinki, Series B: Phonetics, Logopedics and Speech Communication 2, 43—51.

[15] Wiik, K. (1965), *Finnish and English Vowels.* Annales Universitatis Turkuensis. Series B, Tom. 94. Turku: University of Turku.

[16] Söderholm, A-L., Korpijaakko-Huuhka, A-M, Lehtihalmes, M., Juvas, A., Jääskeläinen, T. & Lindqvist, C. (1995), Speech and swallowing defects after oral cancer surgery. *A paper presented at 2nd EORTC International Hong Kong Symposium on Current Trends in Cancer Care*, 13-15 February.

[17] Kent, R.D., Kent, J.F., Weismer, G., Martin, R.E., Sufit, R.L., Brooks, B.R. & Rosenbek, J.C. (1989), Relationship between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. *Clinical Linguistics & Phonetics* 3, 347-358.

ANTICIPATORY MOTOR PROGRAMMING IN ATAXIC DYSARTHRIA

P.F. McCormack & J. C. Ingram~*
**Department of Speech Pathology Flinders University, South Australia, &*
~Department of English, University of Queensland

ABSTRACT

Ataxic dysarthria has prosodic characteristics in which the production of stress and rhythm is disturbed. An experiment is reported on the production of "stress shifts" requiring anticipatory planning by 10 speakers with ataxic dysarthria, compared to 10 matched speakers with normal speech production. Comparison with normal speakers indicated that the ataxics did not take account of the position of the main stress in the following word in their production of "shift" words. This pattern is consistent with a disruption to motor programming.

INTRODUCTION

Ataxic dysarthria arises from damage to the cerebellum or the cerebellar pathways [1]. Its unique prosodic characteristics have often been termed "scanning speech", where each word or syllable appears to be produced with equal prominence, giving the listener the impression that each word is being produced seperately from the others in the utterance. It has been suggested [2] that this prosodic disturbance could arise from either difficulties in the cerebellum's regulation of the ongoing execution of movement sequences, or from impairment to its role in the anticipatory motor programming of such sequences. In order to distinguish between these 2 quite distinct disruptions to speech motor control, an aspect of prosody that is dependent on sequential lookahead for its functioning needs to be investigated.

The stress pattern of some English words in connected speech is dependent on the stress pattern of the following word. It is generally acknowledged that these shifts in the prominence pattern on some words are due to a strong rhythmic constraint to prefer the alternation of stressed and unstressed syllables and words, and to avoid the juxtaposition of stresses [3] [4]. Hence, while *ornate* spoken in a noun phrase such as *the ornate one* has the main stress on the last syllable, in a noun phrase such as *the ornate cup* there is a perception that the main stress has shifted to the first syllable of *ornate*. This rhythm rule can be formulated as an operation where stress shifts from one syllable of a word on to another in order to avoid "clashing" with an adjoining stress.

PROCEDURE

Ten subjects with multiple sclerosis, and who were native speakers of English, were selected. They had been diagnosed by a neurologist as exhibiting predominantly ataxic symptoms, and by an experienced speech pathologist as exhibiting ataxic dysarthria with "scanning speech" prosody. Each subject was recorded reading a series of sentences containing noun phrases which comprised of a potential stress shift word followed by a word with varying syllable distance to its main stress. The sentences were designed to provide a phonological context where shift and non shift environments could be manipulated. Examples of the 5 contexts used are as follows:

Two contexts where no shift was predicted:

No stress following: *There were thirteen of them at the party.*

Shift word focused: *There were THIRTEEN officers at the party.*

Three contexts where shift was predicted:

One syllable distance: *There were thirteen 'officers at the party.*

Two syllable distance: *There were thirteen o'fficials at the party.*

Three syllable distance: *There were thirteen poli'ticians at the party.*

Six potential stress shift words were used: *thirteen*, *bamboo*, *sardine*, *underdone*, *overnight*, and *japanese*. These words had been identified in a previous experiment as being particularly susceptible to stress shift in speech production.

ANALYSIS

Recorded shift words, embedded in their noun phrases, were digitised at 20.8 kHz using the Soundscope speech signal processing program. The duration of the shift word, the duration of each foot, and the duration of the pause between the shift word and the following word was measured. In order to obtain a measure of variation in the duration of the first foot compared to the second foot, the duration of the first foot as a percentage of the duration of the whole word was calculated (relative duration). The peak fundamental frequency for each foot was also calculated using a peak-picking algorithm within the Soundscope program. In order to obtain some measure of the relative changes in fundamental frequency pattern between the 2 feet over different contexts, the value for the second peak was subtracted from that of the first. Three phonetically trained linguists were asked to rate the stress levels in each shift word token as either: 1) the last stressed syllable is more prominent 2) both stressed syllables have equal prominence, or 3) the first stressed syllable is more prominent.

RESULTS

Both the ataxic group and the control group were perceived as shifting stress in the 3 Rhythm contexts, and not shifting in the 2 non rhythm contexts. A 2 way analysis of variance for group membership and context against the judges' perception of stress shift indicated that there were significant main effects for both group and context as well as a significant 2 way interaction between them (context: $p = .000$, $F = 427.3$, d.f = 4, 596; group: $p = .000$, $F = 50.7$, d.f. = 1, 599; interaction: $p = .000$, $F = 24.5$, d.f. 4, 596). The most important difference between the 2 groups was in the pattern of stress shift over the 5 contexts. While the control group displayed a graded decrease in the perception of shift as the syllable distance to the main stress in the fulcrum word increased, the ataxic group showed no such systematic pattern. The ataxic group were perceived as shifting to the same extent in all 3 Rhythm contexts.

There were acoustic-phonetic changes in the shift words that corresponded to these perceived patterns of stress shift in both groups. A 2 way analysis of variance for group membership and context against the relative duration of the first foot in each shift word indicated that there was a significant main effects for context but not for group membership ($p = .000$, $F = 67.1$, d.f. 4, 596). As well there was a significant 2 way interaction between context and group ($p < .05$, $F = 3.4$, d.f. = 4, 596). Both the ataxic and control groups had similar significant increases of approximately 9% in the relative duration of the first foot in the Rhythm contexts compared to the 2 non rhythm contexts. However, the pattern of increase in relative duration was not the same for the 2 groups. The control group displayed a graded decrease in the relative duration of the first foot as

the syllable distance to the main stress in the fulcrum word increased. The ataxic group showed no such systematic difference across the 3 Rhythm contexts, with the degree of durational change remaining the same. Figure 1 displays an error bar plot of the mean relative duration of the first foot across the 5 contexts for the control and ataxic groups.

There were similar results for the shift in peak fundamental frequency between the 2 feet in each word. A 2 way analysis of variance for context and group membership against shift in peak fundamental frequency between the 2 feet indicated significant main effects for both context and group, as well as a significant interaction between them (context: $p = .000$, $F = 57.6$, d.f. = 4, 596; group: $p < .01$, $F = 7.4$, d.f. = 1, 599; interaction: $p = .000$, $F = 11.5$, d.f. = 4, 596). For both the control and ataxic groups there were significant positive shifts in peak fundamental frequency in the Rhythm contexts compared to the non rhythm contexts. The ataxic group differed from the control group, however, in having no graded decrease in peak fundamental frequency shift as the syllable distance increased.

In the ataxic group, as the number of syllables to the main stress in the following word increased, so did the duration of the pause prior to the production of that word. This pattern was markedly different from what occurred in the speech of the control speakers. In the speakers with normal speech production there was no relationship between the metrical structure of the word following the shift word and the pause before the production of that word. A 2 way analysis of variance for context and group membership against pause duration indicated a significant 2 way interaction between them ($p = .000$, $F = 17.5$, d.f. = 4, 596). Figure 2 displays an error bar plot of the mean duration of the pause between the shift and fulcrum words across the 5 contexts for the 2 groups.

It may be thought that the difference in pattern between the ataxic and control groups was a reflection of their differences in rate of speech rather than any impairment to motor control as such. However, a comparison with controls speaking at a slow tempo [5] indicated that the rhythm rule pattern produced by normal speakers at a slow tempo was markedly different from the ataxic productions. The ataxic group were perceived as shifting in all Rhythm contexts, and to the same extent across those contexts. The normal subjects at the slow tempo were not perceived as shifting, nor was there any differential effect in inter-word pausing.

DISCUSSION

The results for the ataxic group indicate that they underwent stress shift in the appropriate contexts, but were unable to make graded phonetic adjustments to the number of syllables to the main stress in the following word. Anticipation of the metrical structure of the following word, however, was found in the duration of the inter-word pause. This pattern was not found amongst the normal speakers, where anticipation of a following word is reflected in the duration of the prior word rather than in the pause [6]. This pattern reflects either a voluntary or involuntary strategy to limit motor programming and production to single words without reference to the following words in an utterance. This suggests that the traditional classification of the speech disorder arising from cerebellar damage as a dysarthria (and therefore not a speech programming disorder) is highly questionable.

REFERENCES

[1] Darley, F., Aronson, A., & Brown, J. (1975). *Motor Speech Disorders. Philadelphia*: Saunders.

[2] Kent, R.D., Netsell, R., & Abbs, J.H. (1979). Acoustic characteristics of dysarthria associated with cerebellar disease. *JSHR, 22,* 627-648.

[3] Liberman, M. & Prince, A.(1977) *On stress and linguistic rhythm,* Linguistic Inquiry, 8, 249-336.

[4] Selkirk, E. (1984) *Phonology and syntax: The relationship between sound and structure.* Camb., Mass.: MIT Press.

[5] McCormack, P. & Ingram, J. (1994). Tempo and the Rhythm Rule. *Proceedings of the 5th Australian International Conf. on Speech Science and Technology.* (pp. 310-315). Perth: UWA Press.

[6] Levelt, W. (1989). *Speaking: from intention to articulation.* Camb., Mass.: MIT Press.

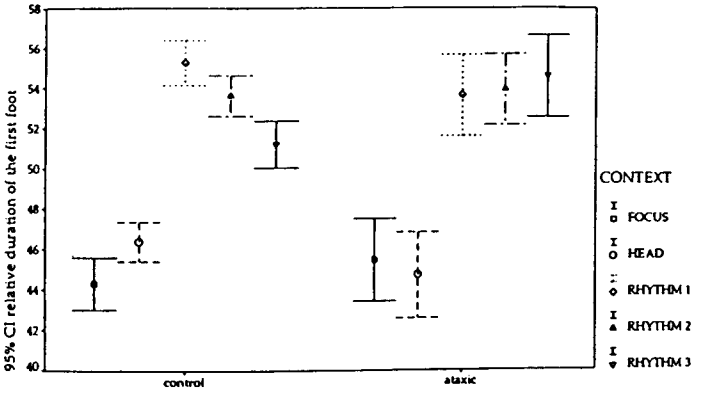


Figure 1 Error bar plot with 95% confidence intervals of the mean relative duration of the first foot across the 5 contexts for the control and ataxic groups (in seconds).

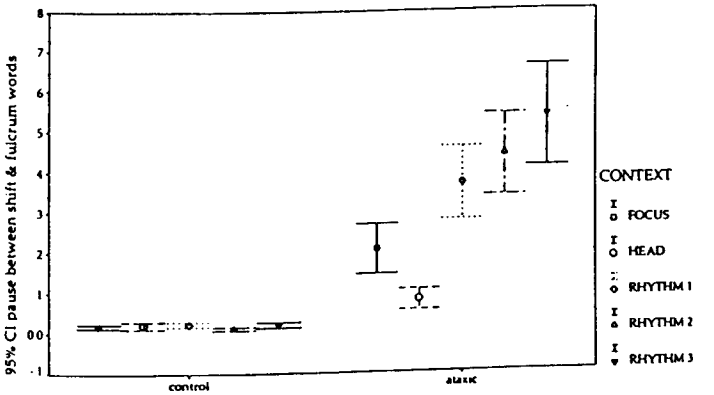


Figure 2. Error bar plot with 95% confidence intervals. Mean duration of the pause between the shift and fulcrum words across the 5 contexts for the control and ataxic groups (in seconds).

A SUPPLEMENTARY TACTILE SPEECH AID TRANSMITTING F0-INFORMATION

Hans Georg Piroth and Thomas Arnhold
Institut für Phonetik und Sprachliche Kommunikation der Universität München, Schellingstr. 3, D-80799 Munich, Germany

ABSTRACT

First experiments are reported on a method for electrotactile F0-coding enabling hearing-impaired persons to detect intonation tactually while using a hearing aid. It is suggested that speech perception can be enhanced by simultaneously delivering segmental information by ear and suprasegmental features by skin.

INTRODUCTION

Within the framework of a TIDE research project on the development of a signal conditioning communication aid for the hearing-impaired [1] it is argued that the selection of the tactile senses as a channel for transmitting suprasegmental features would facilitate the use of residual hearing for the auditory perception of segmental cues.

Early investigations [2] have already shown that frequency differences in a.c.-stimulation can be perceived for frequencies below 200 Hz with just noticeable differences of 1-2 %. Thus, prima facie, electrocutaneous transmission seems to provide one appropriate method for heteromodal presentation of F0-information. But in addition it must be taken into account that identifiability, discriminability and discomfort of electrotactile stimuli are strongly dependent on stimulation methods and electrodes used. Advantages of vibrotactile stimulation have also been discussed [3]. Nevertheless, frequency discrimability is similar for both tactile stimulation modes for low frequencies, and the hardware needs for electrocutaneous stimulators are easier to meet.

APPARATUS AND SUBJECTS

The electrocutaneous stimulation device SEHR-3 [4] used to carry out the experiments produces bipolar rectangular impulses of variable amplitudes and durations which can freely be arranged in sequences. Impulse sequences are delivered to the skin by pairs of circular gold-layered electrodes (9 mm in diameter each with a center-to-center distance of 10 mm between the electrodes of each pair).

Based on pre-test results single impulses had a duration of 208 µs and amplitude was adjusted by subjects before each test session to a subjectively optimal value in the range of 0.33 to 5 mA. Three subjects participated in Exp. 1, seven in Exp. 2.

EXPERIMENT 1

Procedure

To code F0 the method developed by Sparks [5] to construct the apparent movement phenomenon has been adopted. Sparks built patterns of repetitive tactile pulses evoking oscillating movements along a lineal multistimulator array and varied pulse repetition frequency. His results showed that subjects could clearly distinguish successive stimulation (i.e. no apparent movement) at low frequencies, good apparent movement, and partial movement (i.e. a movement perceived along a part of the overall distance between the edge stimulators arranged in line, but of greater subjective strength) at high frequencies. Such a design should be suitable for a tactile transformation of F0 and - if transposed into an appropriate frequency range it also allows marking of out-of-range F0 values by crossing the boundaries of the categories found by Sparks when varying frequencies.

To evaluate the usability of apparent movement for this purpose using the impulse forms and electrode sizes of SEHR-3 two series of electrical pulse sequences were designed. Both series activate adjacent electrode pairs of lineal stimulator matrices successively at constant rate. Series I consists of six pulse sequences with pulse repetition rates of 100, 91, 83, 77, 71, and 67 Hz to cover the transition from partial to good movement. For Series II 43, 37, 32, 29, 26, and 23 Hz were chosen as repetition rates (good to successive movement). Three arrangements of two electrode pairs were used. For each test run, one of these electrode matrices was applied to the dorsal side of the left forearm. The distal electrode pair was placed 2 cm from the wrist, the proximal one 2, 4 or 6 cm apart. Ten repetitions of each pulse sequence were presented in random order. Subjects had to identify whether they sensed successive movement or good apparent movement in one subtest, good or partial in another.

Results and Discussion

Fig. 1 shows that the three categories given could be identified, but that

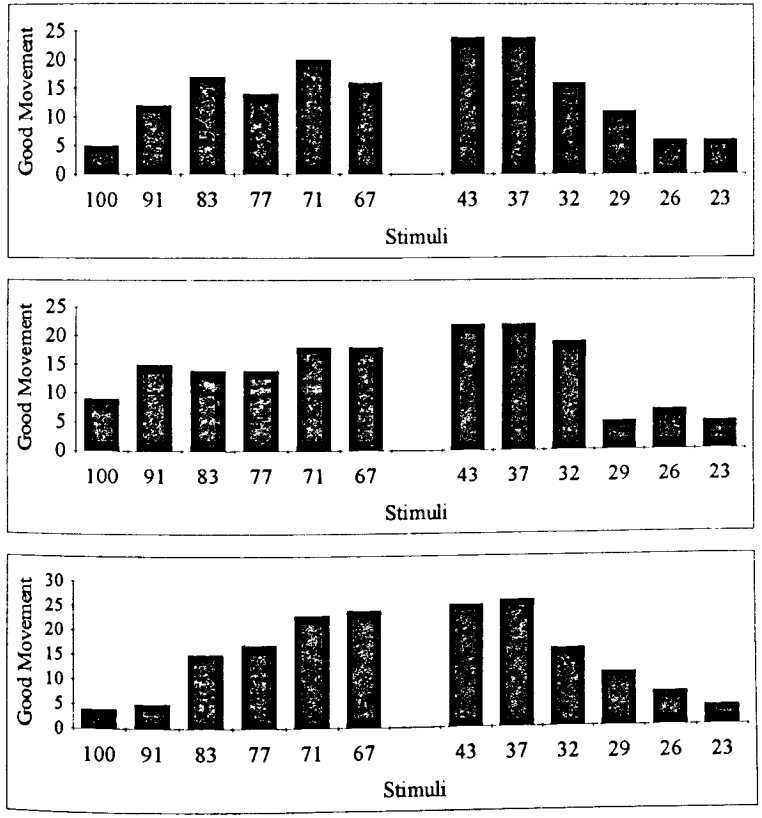


Figure 1. Results of Experiment 1 (Good apparent movement identification). Distance between electrodes: 2 cm (top), 4 cm (mid), and 6 cm (bottom). Scalings in Hz and absolute answers.

recognizability was clearest for the largest distance between electrode pairs. As a conclusion from Exp. 1, this electrode arrangement was chosen for tactile F0 presentation, and it was suggested as a working hypothesis that presenting tactile F0 with a fundamental pulse repetition rate equivalent to the half of an average male speech F0 should be expected to yield optimal results.

EXPERIMENT 2

Procedure

Exp. 2 was designed to test the identification of frequency rises and falls in tactile impulse sequences. For Exp. 2a two series of stimuli were constructed as in Exp. 1, but impulse repetition rate remained constant only for the first 350 ms of the sequence (50 Hz for Series I and 70 Hz for Series II) and continuously changed during the final 250 ms (target rates for Series I: 30, 40, 60 and 70 Hz, for Series II: 50, 60, 80 and 90 Hz). Both series were tested separately (10 repetitions in randomized order) and subjects were asked to identify positive or negative acceleration in the sequences. For Exp. 2b the overall duration of the stimuli was 500 ms, and instead of the final change, frequency peaks with flat or steep slopes were inserted as visualized in Fig. 2. Again, Series I and II were tested separately, but different slopes were mixed (10 repetitions of 4 peak forms with 50% level stimuli added). Subjects had to answer whether they perceived a frequency rise or not.

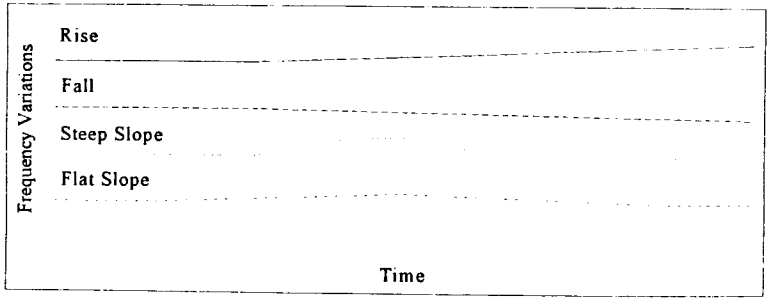


Figure 2. Stimulus examples for Experiment 2.

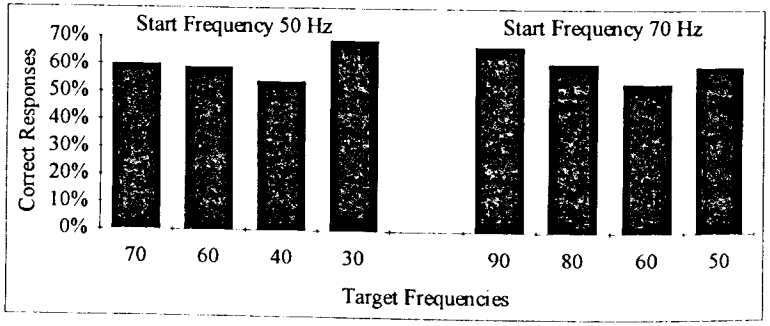


Figure 3a. Results of Experiment 2a (Identification of Rises and Falls).

Results and Discussion

Fig. 3a shows correctly identified rises and falls in Exp. 2a. Greater frequency changes yield better results in either direction. The highest rates are found for the 70-90 Hz rise and the 50-30 Hz fall

which both reach the boundary of the good apparent movement category, but overall recognizability of rises and falls is less than expected.

Fig. 3b, presenting the results of Exp. 2b, reveals one hint to the explanation of these low rates: identification of F0 peaks is similar to the results of Exp. 3a, but also in about 60% of the stimuli with constant F0, frequency changes were perceived.

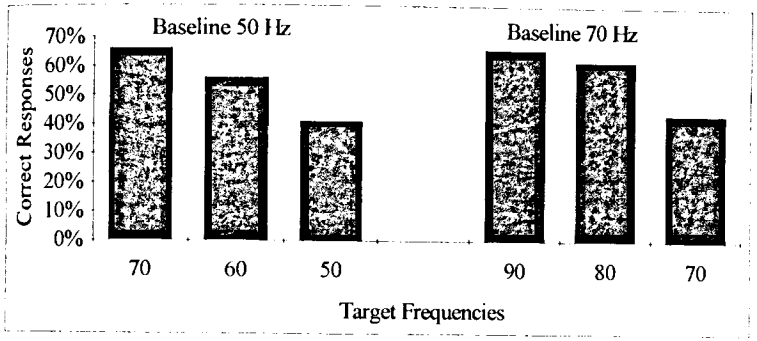


Figure 3b. Results of Experiment 2b (Identification of Peaks and Levels).

GENERAL DISCUSSION

The experiments show that the categorical aspects of the apparent movement phenomenon were clearly recognized, but that a transposition of speech F0 into the good apparent movement range alone is not sufficient to code electrotactile F0. Since subjects often perceive F0 peaks in stimuli with constant frequency, it can be suggested that frequency must decrease over the pattern to ensure the perception of a clearly constant frequency. This would mean that speech F0 declination must be preserved in the transposition method for tactile stimulation and, since the identification of tactile frequency shifts was poorer than expected, F0 variations have to be increased.

ACKNOWLEDGEMENT

Research was supported by the Commission of the European Communities (TIDE project No. 1090: SICONA).

REFERENCES

[1] Bauer, D., Geiger, J. C. & Beerwerth, R. (1993), Speech Signal Conditioning Communication Aids for the Impaired with Severe Auditory Sensory Damages, in: *Speech and Language Technology for Disabled Persons*, ed. by B. Granström, Sh. Hunnicutt & K. E. Spens, Stockholm: ESCA, pp. 51-54.

[2] Monjé, M. (1936), "Über die Wirkung von Wechselströmen verschiedener Frequenz auf die Hautsensibilität", *Zeitschrift für Sinnesphysiologie*, vol. 67, pp. 2-18.

[3] Summers, I. R., Gratton, D. A., Dixon, P. R., Brown, B. H. & Stevens, J. C. (1992), Comparison of Vibrotactile and Electrotactile Modalities with a Variety of Coding Strategies, in: *Proceedings of the Second International Conference on Tactile Aids, Hearing Aids and Cochlear Implants*, ed. by A. Risberg, S. Felicetti, G. Plant and K. E. Spens, Stockholm: Dept. of Speech Communication & Music Acoustics, pp. 33-38.

[4] Piroth, H. G. & Tillmann, H. G. (1991), "Das System zur elektrischen Hautreizung SEHR-3", *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, vol.29, pp. 41-46.

[5] Sparks, D. W. (1979), "The Identification of the Direction of Electrocutaneous Stimulation Along Lineal Multistimulator Arrays", *Perception & Psychophysics*, vol. 25, pp. 80-87.

VOICING, FUNDAMENTAL FREQUENCY, AMPLITUDE ENVELOPE AND VOICELESS-NESS AS CUES TO CONSONANT IDENTITY

Stuart Rosen, Andrew Faulkner, Kirsti Reeve* and Kerensa Smith**
Northwestern University, Evanston, Illinois, U.S.A.
**Department of Phonetics and Linguistics, University College London, U.K.*

ABSTRACT

Of long-standing theoretical and practical interest is the extent to which cues to consonant identity can be provided by purely temporal auditory features (periodic and aperiodic excitation, and amplitude envelope). Here we show that the primary features used by normal observers involve the on-and-off patterning of silence, periodicity and aperiodicity (both with and without lipreading). Additional variations in envelope and fundamental frequency provide little further information.

INTRODUCTION

For some years, we have been developing a speech-pattern hearing aid for profoundly hearing-impaired people [1]. The original SiVo aid (**Si**nusoidal **Vo**ice) extracted from the speech only the voice fundamental frequency (Fx) and presented it as a sinusoid at a constant loudness — a signal which provides an effective auditory supplement to lipreading. Now, further speech pattern elements are being incorporated, representing the speech amplitude envelope and voiceless frication, and we wanted to obtain normative data for comparison with results from our hearing-impaired listeners. Not the least reason for this is practical — to know if our current tests would be sensitive to these extra acoustic features.

Having decided to focus on the perception of intervocalic consonants for the moment, there were other interesting issues to address. For example, it has often been noted that Fx contours have a microstructure that could, in theory, transmit segmental information over and above that contained in the simple on-and-off pattern of voicing.

There is also currently much interest in the temporal structure of speech [2, 3] and in particular, the degree to which amplitude envelope is important. At least one source of this interest is the extent to which amplitude compression in auditory prostheses, with its transformation of the natural envelope of speech, would have a deleterious effect. Here we compare the most extreme compression (signals with no variation in amplitude when "on") to signals with natural variations in envelope.

EXPERIMENT 1

Experiment 1 investigated a simple coding of the voiced components of speech only. Represented were the on-off pattern of larynx excitation, its' fundamental frequency, and the amplitude envelope. The key questions were whether Fx variation and/or envelope provide cues to consonant identity beyond those in the on-off pattern of larynx excitation.

Methods

A total of 9 conditions were tested: lipreading alone (**L**), plus 4 sound conditions with (**L+**) and without lipreading:

V - A fixed-frequency, fixed-amplitude signal indicating vocal fold vibration.

V(A) - as for **V** but with added amplitude envelope, derived from the original speech.

Fx - A fixed-amplitude signal whose periodicity followed the speaker's Fx.

Fx(A) - as for **Fx**, but with an amplitude envelope added.

Four normal-hearing native speakers of British English took part. Speech materials comprised each of the 24 English consonants between the vowel /ɑ/. Five distinct video-recorded lists (female speaker) were employed, each consisting of 2 tokens of each consonant. One list was reserved for initial training in each condition, whilst 4 were used for testing.

Fundamental frequency and voicing information were recorded by means of an electro-laryngograph on the speaker's throat at the time of recording, in the form of narrow pulses synchronised to the speaker's vocal fold closures. These pulses were then used as input to an external device with two modes of operation for generating the test sounds. For conditions involving **Fx**, the original pulses were used to trigger other pulses on a 1-for-1 basis. For conditions involving **V**, the original pulses were used to gate on and off a train of pulses of constant frequency. The triggered pulses were low-pass filtered at 400 Hz (18 dB/octave) to make them pleasant to listen to. For conditions with amplitude information, envelopes were derived by full-wave rectifying the 3-kHz low-pass filtered speech, and smoothing the result with a 30 Hz low-pass filter. These were then multiplied against the appropriate pulse train. All signals were recorded for testing purposes, and presented free-field using a loudspeaker.

Analysis

Each session was analyzed separately by constructing a confusion matrix from which overall proportion correct scores were derived, together with unconditional information transfer measures for:

voicing: voiced vs. voiceless

place: bilabial vs. labiodental vs. dental vs. alveolar vs. palatal vs. velar vs. pharyngeal

manner: plosive vs. affricate vs. fricative vs. nasal vs. glide

voice/man: a slightly collapsed voicing/manner feature, closely related to so-called envelope features [3] — voiced plosives vs. voiceless plosives vs. voiceless fricatives vs. sonorants (nasals + glides) vs. voiced fricatives.

To allow for learning, only the last 6 sessions for each condition of the 10 run were analyzed. Statistical claims are made on the basis of an ANOVA including an observer x condition interaction (which was often significant), and Tukey's Studentized Range Test ($p \leq 0.05$).

Results

Table 1 shows mean performance as a function of condition. Values with a common symbol in the same column (*, #, @) are indistinguishable statistically. Although more information tends to lead to better performance, neither fundamental frequency nor envelope increase performance very much compared to on-off voicing. That Fx variations aid consonant identification little has already been shown [4], but the small effects of envelope variation come as a surprise.

TABLE 1

condition	*feature*				
	correct	*voice/man*	*voicing*	*manner*	*place*
V	# 13	# @ 45	@ 68	# 28	# 22
V(A)	# 14	# @ 48	@ 69	# 30	# 23
Fx	# 18	@ 52	@ 75	# 35	# 25
Fx(A)	# 17	# @ 50	@ 72	# 33	# 24
L	54	# 43	15	60	* 90
L + V	* 79	* 71	* 92	@ 67	* 93
L + V(A)	* 83	* 76	* 93	@ * 73	* 95
L + Fx	* 83	* 77	* 94	* 75	* 93
L + Fx(A)	* 85	* 78	* 95	* 77	* 94

EXPERIMENT 2

Experiment 2 was primarily concerned with the role of voiceless frication and envelope. There were 3 different sound signals, presented with and without lipreading, making for a total of 6 conditions. Apart from **Fx(A)** used in Experiment 1, the other sound signals were:

Fx(A)+Nz - as for **Fx(A)** above, with a band of fixed-level noise present during periods of voiceless excitation.

Fx(A)+Nz(A) - as above, but with an amplitude envelope on the noise as well.

Methods

Five new observers took part, following the same procedure as in Experiment 1. Signal components reflecting voicing in the speech signal were created as described for Experiment 1. Voiceless excitation was detected by a spectral balance circuit comparing the amount of energy above and below 3 kHz in the speech signal. However, the presence of voice pulses from the laryngograph overrode the comparator. Thus, voiceless excitation could only be detected in the absence of voicing. When the comparator indicated voiceless excitation, it gated a white noise that was then mixed with the voicing pulses, for final low-pass filtering at 400 Hz. For conditions with amplitude information, envelopes were derived by full-wave rectifying the broad-band speech signal and smoothing the result using a 30 Hz low-pass filter. These envelopes were then multiplied against the white noise. Both the noise and pulse train signals were low-pass filtered at 400 Hz. Again, all signals were recorded for testing purposes.

Results

Analysis procedures were the same as described for Experiment 1, resulting in the summary found in Table 2. Again, more information tends to lead to better performance. The addition of voiceless information almost always leads to significantly improved performance (except for the place feature). However, the addition of amplitude envelope never causes significant increments for the features analysed (just as found in Experiment 1).

TABLE 2 condition	*feature*				
	correct	*voice/man*	*voicing*	*manner*	*place*
Fx(A)	# 19	# 47	62	# 32	# 24
Fx(A) + Nz	@ 24	@ 60	# 72	@ 45	# 24
Fx(A) + Nz(A)	@ 27	@ 61	@ # 79	@ 45	# 27
L + Fx(A)	68	68	@ 82	61	* 79
L + Fx(A) +Nz	* 76	* 82	* 92	* 73	* 80
L + Fx(A) +Nz(A)	* 75	* 79	@ * 88	* 72	* 82

EXPERIMENT 3

Experiment 3 focused primarily on the overall role of envelope, using the same methods as previously, but with three new observers. Seven conditions were used, involving lipreading alone, plus three sound signals both with and without lipreading: **Fx**, **Fx+Nz** (which had not been used previously), and **Fx(A)+Nz(A)**.

Results

The results (Table 3), lead to essentially the same conclusions as the previous two experiments. Variations in envelope beyond a simple binary indication of amplitude never lead to statistically significant increments in performance. But the addition of voiceless information often does, especially for voicing and for other features in conjunction with lipreading.

TABLE 3 condition	*feature*				
	correct	*voice/man*	*voicing*	*manner*	*place*
Fx	# 9	# 33	# 33	# 26	# 21
Fx + Nz	# 14	# 39	@ 50	# 28	# 22
Fx(A) + Nz(A)	# 14	# 39	@ 48	# 28	# 22
L	@ 41	# 33	5	@ 48	@ 77
L + Fx	62	58	@ 60	56	@ * 79
L + Fx +Nz	* 72	* 69	* 73	* 67	@ * 81
L + Fx(A) +Nz(A)	* 74	* 73	* 75	* 70	* 82

DISCUSSION

At first sight, these results bode well for auditory prostheses that distort envelope. Insofar as envelope variations made little difference to performance, it is clear that the bulk of temporal segmental information is contained in the on-and-off patterning of silence, periodicity and aperiodicity.

But there are two important caveats. First, it is not possible to extrapolate to connected discourse from results with consonants. We already know that variations in Fx aid consonant identification little in comparison to a simple voicing indicator, even though such variations are of great utility in connected discourse [5]. And there is evidence that envelope variations are as much a benefit in connected discourse as are Fx variations [6]. Second, it may not be wise to extrapolate from the results of normal observers to impaired ones. Faulkner *et al.* [1] have already shown that for some profoundly hearing-impaired observers listening to signals analogous to those used above, the addition of envelope can be of significant benefit, both in connected discourse *and* in consonant identification. It may be that impaired listeners are less able than our normally hearing listeners to use the on-off timing of voiced and voiceless excitations, and in consequence, may depend more on the use of other correlated cues conveyed by amplitude envelope.

ACKNOWLEDGEMENTS

Supported primarily by the Medical Research Council (UK), with important additional support from a Wellcome Trust Vacation Scholarship, CEC TIDE projects 206 (STRIDE) and 1217 (OSCAR), and Northwestern University.

REFERENCES

[1] Faulkner A *et al.* (1992), "Speech pattern hearing aids for the profoundly hearing-impaired: Speech perception and auditory abilities", *J Acoust Soc Am*, vol. 91, pp 2136-2155.

[2] Rosen S (1992), "Temporal information in speech: acoustic, auditory and linguistic aspects", *Phil Trans Royal Soc London B*, vol. 336, pp 367-373.

[3] Van Tasell DJ *et al.* (1992), "Temporal cues for consonant recognition: Training, talker generalization, and use in the evaluation of cochlear implants", *J Acoust Soc Am*, vol. 92, pp 1247-1257.

[4] Rosen S *et al.* (1979), "Lipreading with fundamental frequency information", *Proc Inst Acoust Autumn Conf*, pp 5-8.

[5] Rosen S *et al.* (1980), "Lipreading connected discourse with fundamental frequency information", *Brit Soc Audiol Newsletter (Summer)*, pp 42-43.

[6] Grant KW *et al.* (1985), "The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects", *J Acoust Soc Am*, vol. 77, pp 671-677.

PROSTHETIC REMEDIATION OF SPEECH QUALITY FOR POST VELUM RADIOTHERAPY USING PHONETIC INVESTIGATION METHOD

*Bernard Teston * and André Besson ***
**Laboratoire Parole et Langage CNRS, Aix-en-Provence, France*
*** Hôpital Paoli-Calmettes CRAM, Marseille, France*

ABSTRACT

Patients having undergone irradiation in the pharyngeal area often exhibit a lingual-palatal articulatory deficiency in addition to the velar disorder. This paper shows how a simple but carefully made palate cover plate based on phonetic principles can alleviate some of the velar defects by decreasing the severity of the rhinolalia and correcting the tongue-to-palate contacts required for correct articulation.

INTRODUCTION

The velum of patients having undergone irradiation in the pharyngeal area often has a normal morphology but remains immobilized in the lowered position. The state of the tissues is generally too poor to allow for restorative pharyngovelar surgery. The only remaining remedy for the velar deficit is an artificial palate. However, such a prothesis is difficult to achieve due to the shape of the velum, which excludes the use of Schiltsky-type obturators. Our aim here is to show that, in such cases, a phonetic approach based on the techniques used to investigate speech production, associated with particular care in making the prothesis, can substantially improve the patient's speech.

1. CLINICAL DESCRIPTION OF A CASE

The patient is a 33-year-old woman. At the age of 17, she was diagnosed to suffer from epidermoid carcinoma of the cavum, accompanied by substantial ganglionic extension. Given the severity of the case, chemotherapy was chosen, in conjunction with massive radiation encompassing the entire cavum.

Today the patient is considered to be cured, although the aftereffects of the radiation therapy are great. The main characteristic of the sequels is an overall stiffness of the musculature in the pharyngeal region. The patient reports that her mouth and pharynx feel as if they "form a single block."

In its atonic state, the velum remains lowered. The upper tongue muscle is altered in an asymmetrical manner, which leads to substantial deviation to the right and a twisted tongue. The left dorsal surface also exhibits a concave type of dysmorphology. The patient's jaw opening at the central incisors is reduced to 18 mm. The patient also suffers from partial asialia and a slight hindrance in swallowing. Her speech is highly affected, with significant rhinolalia accompanied by difficult lingual-alveolar articulation, worsened by incorrect lip dynamics which tend to be used as an offset for her overall pronunciation problems. Her difficulty speaking, broken by frequent breath taking, is a real handicap with which the patient has trouble coping (she is a psychologist in an unemployment office).

As pharyngovelar restorative surgery is not advisable, the maxillofacial surgery ward of the Paoli-Calmettes Hospital referred the patient to our laboratory for potential fitting with a prothesis.

2. PHONETIC DESCRIPTION OF THE CASE

2.1 Assessment of Rhinolalia

The first step was to determine the severity of the patient's rhinolalia and the residual motricity of the velum. This was done using an aerophonometer, a device which provides accurate measurements of the buccal and nasal air flow rates [1]. The aerophonometer was developed for this purpose as part of the EVA system [2]. The evaluation corpus consisted of two sentences which the patient was asked to pronounce several times: "ta toupie va trop vite" /tatupivatrovit/, designed to detect nasal leakage on the consonants "p" and "t", and "ta tante a chanté" /tatãtaʃãte/, aimed at displaying the movements of the velum via variations in the nasal air flow rate on the

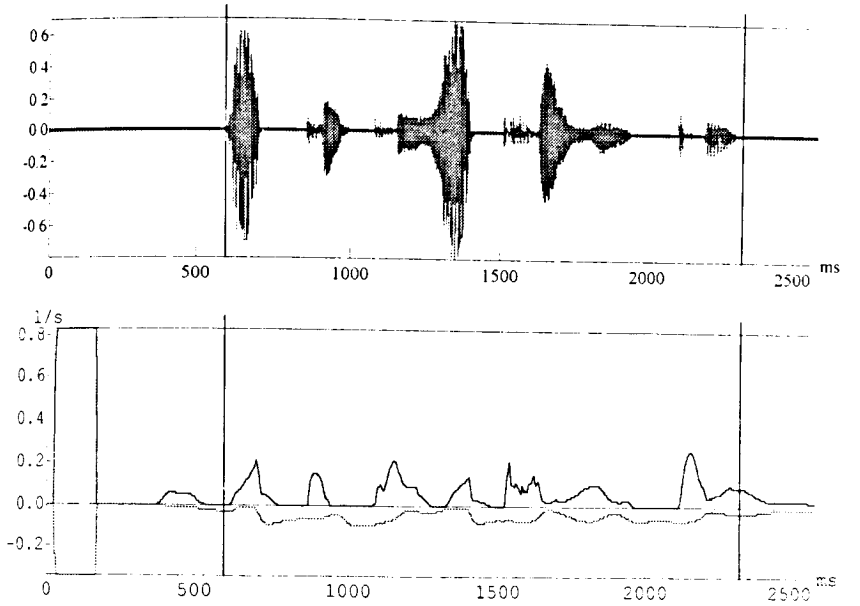


Figure 1: " latupivatrovit *" before prothesis. Upper trace: Spl. Lower trace: oral airflow and inversed nasal airflow*

For the ten repetitions of each of the two sentences, the total volume of air exhaled from the nose and mouth were calculated, in addition to the ratio of the nasal air volume to the total volume exhaled. For the first sentence, an average of 51% of the air was exhaled through the nose (for a normal subject, this average is less a few percents).

Figure 1 shows the closest production to the mean for the first sentence. We can see a heavy and nearly constant nasal leakage everywhere except on the nuclei of the vowels "a" and "o", with the maximum occurring on the voiced consonant "v" and the stops "p" and "t". The movement span of the apical-alveolar and labial apparatus is small, the bursts of the consonants "p" and "t" are not very prominent, and the acoustic signal on the voiced consonant "v" is weak. This indicates atonic articulation: the subject "holds back" in order to avoid having to take too many breaths. For the second sentence, nasal leakage is approximately the same (49%) on the closure of the consonants "t" and "ch" in the nasalized environment. The airflow rates for the oral and nasal vowels are the same, which confirms the fact that the velum remains immobile and lowered during phonation.

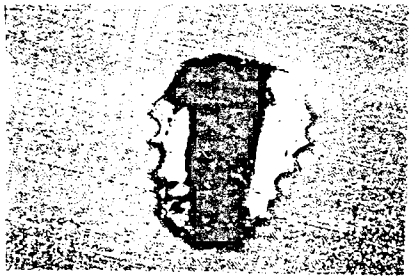
2.2 Palatographic examination

The tongue asymmetry observed clinically was confirmed by palatography [4].

As it was impossible for the subject to open her mouth so that we could photograph her palate, we used an alternative approach which consisted of coating a palate plate and photographing the tongue after articulation. The purpose of this test was to point out the areas of excessive and insufficient contact brought about by the lingual asymmetry.

The consonants tested were "t", "s", and "k" surrounded by the vowel "a". For "ata", the tongue made a horseshoe type of contact, with a narrower area on the left. For "asa", there were two lateral contacts, again with a smaller area on the left. For "aka", the contact was not entirely closed due to the lack of left paramedial contact (figure 2).

[asa]



[asa]

Figure 2 : asymetry of the lingopalatal contacts

Although this investigation was skewed to some extent by the change in procedure, it nevertheless allowed us to determine what stop consonant compensations were occurring on the inner arch of the palatal plate.

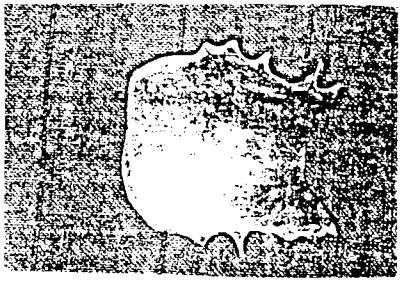
3. MAKING THE PALATOVELAR PROTHESIS

The patient's normal but rigid velar morphology prevents the insertion behind the velum of a support for a Schiltsky type of pharyngeal obturator. For cases like this, we designed a simple palatovelar cover plate capable of raising the velum and thereby reducing the severity of the rhinolalia. Such a plate should also allow the subject to more evenly distribute tongue contacts in the areas necessary for correct articulation.

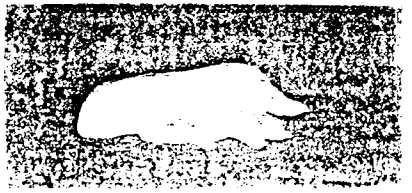
Due to the patient's narrow buccal opening, obtaining an imprint of the palate posed an additional problem which was solved by making a custom-designed imprint carrier. Several imprints were made to ensure the quality of the mold, which was then used to make the palatographic plate and several palatal plates in biocryl resin.

The reference plate is 8/10 mm thick, with a sagittal profile that extends the hard palate. Its posterior edge is located 4 mm from the uvula. On this basic plate, a thickened area on the left inner arch is used to correct the defective occlusion caused by the lingual asymmetry. The thickness needed for correction was tested empirically, under continuous palatographic control, until the best trade-off was obtained between "too much and not enough" occlusion.

Three plates were produced from the reference plate (figure 3), each with a gradual increase in the thickness of the arch starting at the beginning of the postdam and ending at an added thickness of 2 mm, 4 mm, and 6 mm, respectively.



inner velopalatal plate



profile velopalatal plate

Figure 3: velopalatal prothesis

4. RESULTS OF THE REHABILITATION

The patient felt "supported" and relieved during articulation as soon as she began wearing the 2-mm plate, which she perceived as a prop during continuous speech efforts. Habituation to this 2-mm plate took approximately one month. The 4-mm plate was also worn for one month, after which discomfort was eliminated.

Figure 4 shows the mean nasal air volume of 32% achieved on the first sentence. This is not a very large gain but allows the subject to take breaths less frequently and to achieve better apical-dental and labial articulation. Stop consonants such as "p" and "t" are more clearly marked, and the signal on the voiced consonant "v" is not as weak. According to individuals in the subject's surroundings, the patient's speech is more natural and her voice has a more pleasant quality. However, in exchange, the subject can only swallow liquids.

After these first two months, the subject was fitted with the 6-mm plate. She is now capable of making well-articulated speeches in public, without excessive fatigue. A final metal plate was made.

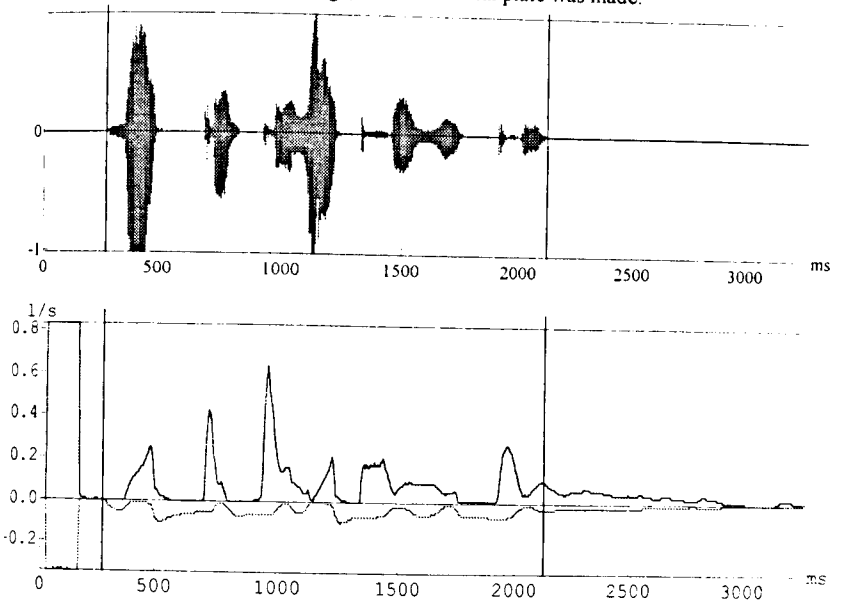


Figure 4: " latupivatrovit " with 4 mm plate prothesis. Upper trace: Spl. Lower trace: oral airflow and inversed nasal airflow

CONCLUSION

Designing and developing such a prothesis requires sustained efforts on the part of the patient, and careful work and substantial knowledge about the physiological mechanisms of speech production on the part of the specialist. However, the actual fabrication and use of the prothesis are relatively simple and non-traumatic. In the light of these positive results, we consider this method to be an effective solution to speech rehabilitation in difficult cases where restorative surgery is not recommended. The method has been used on other less or equaly severe cases, and the results have been just as good.

REFERENCES

[1] - Teston B. (1983), "A system for the analysis of the aerodynamic parameters of speech", *10th International Congress of Phonetic Sciences*, Utrecht, Section 5, 457.

[2] - Teston B. et Galindo B. (1995), "A diagnostic and rehabilitation aid work station for speech and voice pathologies", *Proceedings of Eurospeech 95*, 4 pp.

[3]Edwards, M. and Watson A.C.H. (1980) *Advances in the management of cleft palate*,Churchill Livingstone, Edinburgh, chap. 16, 232-246.

[4] Besson, A. (1974), *Incidence de la téléradiographie et de la palatographie en prothèse vélopalatine*, Thèsis Medical University, Marseille,264.

VOCAL AND SUBVOCAL SPEECH IN STUTTERERS AND NON-STUTTERERS

J. van Rie, Dept. of General Linguistics and Dialectology
A.C.M. Rietveld, Dept. of Language and Speech
University of Nijmegen, The Netherlands

ABSTRACT

The methodology of subvocal speech was used to assess whether stuttering already exists at the pre-motor stage of the speech production process. In realizing CVCV sequences non-stutterers (NST) were faster than stutterers (ST), both in vocal and subvocal speech. Also, ST produced vocal speech just as fast as subvocal speech, what leads us to conclude that in subvocal speech of ST different factors play a role than in subvocal speech of NST.

INTRODUCTION

The deviant speech of ST has been approached from many points of view. Differences with speech of NST have been searched for in linguistic planning [1], in the articulatory planning and in articulatory execution of speech. Especially in the last two approaches the concept of *articulation difficulty* plays an important role. Specific segments, like initial /g,d,l,p/, are assumed to be more difficult for ST than others, like /w,s,f,h/ [2]. Another factor appears to be the similarity of consonants on identical syllable positions [2]; it was found that consonants which differ by only one Distictive Feature (DF) enhance stuttering, compared to more dissimilar consonants. It is still not clear, however, whether articulatory problems should be located only at the execution level of speech, or also in the planning stage. There is evidence, yet, that articulatory obstacles are to be found in the planning stage as well. [3] for instance, showed for normal speakers a qualitative similarity between 'slips of the tongue' in overt and covert speech, whereas [4] reported that subjects need more time to silently read sentences with tongue twisters than matched sentences without this kind of obstacles. These findings suggest that articulatory problems also show up at the planning stage of speech. The methodology used in the research reported above is that of *silent reading*, which is equivalent to *subvocal* or *covert* speech. It offers the opportunity to tap the speech production process at the pre-execution stage, where movements of the speech organs are not yet initiated, and do not provide feedback in order to signal whether targets are reached or not. This is a particularly favourable situation to assess whether differences between the speech of ST and NST mainly exist at the execution stage, or already at the pre-motor stage.

This contribution focusses on the differences between vocal and subvocal speech of ST and NST. We did not investigate stuttered speech, but restricted ourselves to *perceptually fluent speech*, i.e. speech which is fluent in the overt condition (OC). Matched speech samples in the covert condition (CC) are considered to be fluent too. It is well-known, however, that perceptually fluent speech of ST is often slower than that of NST. Thus the deviance of planning and/or execution in ST' speech can manifest itself in the rate of speech.

Our hypothesis is that stuttering, or its manifestation in fluent speech: lower speech rate, is not (only) located in the articulatory/motoric execution stage. This hypothesis implies the following predictions when comparing vocal and subvocal speech of ST and NST:

1) ST need more time for the realization of speech in both the CC and OC than NST;
2) ST and NST need more time in the OC than in the CC, as the former is an additional, time-consuming part in the process of speech production;
3) The difference in speech durations between ST and NST is less for *simple* sequences than for *difficult* ones.

METHOD

Speech materials

It was decided to use CVCV nonsense words for the sequences to be realized both in OC and CC. There were three reasons why this type of words was used: a) nonsense words leave more freedom for phonetic composition, b) an emotional load for stutterers is avoided, and c) less stuttering is observed on nonsense than on normal words [2]. The words were varied along a number of dimensions, which are assumed to be related to rate of speech or the facilitation of stuttering. [5] found that a relatively large dissimilarity between consonants on corresponding syllable positions increases the rate of speech, while Soderberg [2] observed a high frequency of stuttering on words in which the corresponding consonants differ by only one DF. Thus we created a dichotomy of words which are or are not assumed to stimulate stuttering or reduce speech rate. The dimensions are:

1) The number of DF's in which consonants with corresponding syllable positions differ (0 - 6 DF's). For instance the consonants of the word *piepe* do not differ in their segmental make-up, whereas the word *siene* has consonants that differ in 5 DF's.
2) Initial consonants which are known to facilitate stuttering, like /g,d,l,p/ and consonants which do not facilitate stuttering: /w,s,f,h/.

By combining these two dimensions we eventually tested two types of sequences: *difficult* sequences that maximally enhance stuttering and *simple* sequences that do not facilitate stuttering.

Subjects

Both groups contained 12 subjects matched for sex and age. Both ST and NST were classified according to quantitative stuttering severity by means of the *Stuttering Severity Instrument* [6]. The ST showed *very mild* (N=3), *mild* (N=4), *moderate* (N=4) and *severe* (N=1) stuttering behaviour. The speech of NST was classified as *very mild*, even though most of them showed no dysfluencies at all. A model for a differential diagnosis and treatment of stuttering [7] was used to determine the qualitative stuttering behaviour: in all ST a motor dysfunction was dominant, characterised by lengthening, blocks and non-verbal struggle behaviour.

Procedure

The condition for realizing a sequence was displayed on a computer screen, being either *aloud* (OC) or *quietly* (CC). The subjects was told that *quietly* was the equivalent of repeating a telephone number or a list of shoppings in their mind. During this display the subjects would prepare themselves to overtly or covertly producing the sequence by either opening the lips a little so that they could start articulating the sequence as soon as it appeared on the screen (OC) or by clamping the tongue between the teeth and keeping the lips apart (CC) to prevent the articulators from making the same articulatory gestures as in OC. When the sequence appeared on the screen the subjects had to start repeating it in the proper condition as fast as possible, meanwhile maintaining a precise articulation. With every repetition (both in OC and CC) they simultaneously had to press a key which was connected to a

computer that calculated the average realisation duration of every sequence. This way, in OC, the speech production process at the execution stage was tapped. The durations tapped in the CC represented the execution duration in the pre-execution stage only. In order not to include possible speeding up at the beginning of repeating a sequence and/or slowing down at the end, only the five intervals in the middle were used for computing the average duration of every sequence.

RESULTS AND CONCLUSIONS

The mean realization durations of NST ranged from 217 ms to 554 ms in OC and from 178 ms to 435 ms in CC. For ST the mean durations ranged from 191 ms to 655 ms in OC and from 230 ms to 818 ms in CC. Analysis of variance showed a significant effect for Group ($df=1$, $F(1,22)=4.85$, $p=.04$) and for Difficult versus Simple sequences ($df=1$, $F(1,22)=7.79$, $p=.01$). The interaction between Group, Condition and Difficult versus Simple sequences ($df=1$, $F(1,22)=6,43$, $p=.02$) is depicted in figures 1 and 2.

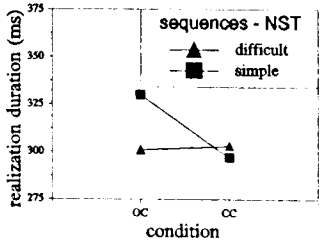


Figure 1. Interaction for NST between Condition and Difficult versus Simple sequences.

Looking at the first prediction made in the introduction, it appears that NST realize all difficult and simple sequences in both conditions faster than ST, the mean difference in CC (69.1 ms)

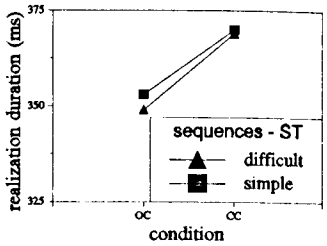


Figure 2. Interaction for ST between Condition and Difficult versus Simple sequences.

being considerably larger than in OC (35.5 ms). That ST would perform the task at a lower rate than NST in OC was to be expected. It was observed in earlier experiments that, in addition to a more general motor slowness [8], ST have an articulatory slowness as a result of increased muscle tension. Even in perceptually fluent speech they significantly differ from NST with respect to speaking rate [9] [10] or physiological characteristics like subvocal pressure [11]. The finding that ST were also slower in CC confirms our first hypothesis that the delay in speech of ST can be reduced to the pre-execution stage. However, we have to be careful with this conclusion as will be seen when discussing the second prediction.

Looking at the overt - covert dichotomy, it appears that NST are slower in OC than in CC, the mean difference being 15.3 ms. As opposed to NST, ST produce the sequences faster in OC than in CC, the difference being 18.3 ms. Although the execution stage is eliminated in CC, ST need more time in CC than in OC. Probably additional factors play a role here, like auditory or proprioceptive feedback or behavioral factors, that could have caused time delay in stead of gain. ST could have

been more aware of their speech than NST. In performing the, very unusual, CC task they might have felt insecure about their performance. In trying to perform as good as possible, ST may have adjusted to the task by taking more time to be able to 'control' their responses, especially when both auditory and proprioceptive feedback were not available as monitoring mechanisms. The influence of sensory feedback is considered to be unequal for the speech of ST and NST, but disordered sensory feedback as an explanation for stuttering is still under debate [12].

The simple versus difficult dichotomy does not seem to influence the realization rate of ST, independent of condition. For NST the same holds for CC, but in OC something strange happens: NST need more time for the sequences that were considered *simple* than for the sequences that were considered *difficult*.

ACKNOWLEDGEMENT

We express our gratitude to M.C. Franken for her support and critisism during the research.

REFERENCES

[1] Koopmans, M.L., Slis, I.H. & Rietveld, A.C.M. (1991), "Stuttering as indication of speech planning", *Proceedings XIIth ICPhS*, Aix-en-Provence, vol. 2, pp. 30-33.

[2] St. Louis, K.O. (1979), "Linguistic and motor aspects of stuttering". In: *Speech and Language*, vol. 1, (N. Lass, ed.).

[3] Dell, G.S. & Repka, R.J. (1992), "Errors in inner speech". In: *Experimental slips and human error: exploring the architecture of violation* (B.J. Baars, ed.), pp. 237-263. New York: Plenum Press.

[4] Haber, L.R. & Haber, R.N. (1982), "Does silent reading involve articulation? Evidence from tongue twisters", *American J. of Psychology*, vol. 95, pp. 409-419.

[5] Smith, B.L., Hillenbrand, J., Wasowicz, J. & Preston, J. (1986), "Durational characteristics of vocal and subvocal speech: implications concerning phonological organization and articulatory difficulty", *J. of Phonetics*, vol. 14, pp. 265-281.

[6] Riley, G.D. (1980), *Stuttering Severity Instrument. For children and adults.* Revised edition. Austin, Texas: Pro-ed.

[7] Kraaimaat, F. & Janssen, P. (1983) "Een werkmodel voor een differentiële diagnostiek en behandeling van stotteren", *Gedragstherapie*, vol. 16, pp. 299-309.

[8] Starkweather, C.W. (1987) "Laryngeal and articulatory behavior in stuttering: past and future". In: *Speech motor dynamics in stuttering* (H.F.M. Peters & W. Hulstijn, eds.), pp. 3-18. Wien: Springer Verlag.

[9] Zimmerman, G.N. (1980), "Articulatory dynamics of fluent utterances of stutterers and nonstutterers", *J. of Speech and Hearing Research*, vol. 23, pp. 95-107.

[10] Starkweather, C.W. & Meyers, M. (1979), "The duration of subsegments within the intervocalic intervals of stutterers and nonstutterers", *J. of Fluency Disorders*, vol. 4, pp. 205-214.

[11] Peters, H.F.M. & Boves, L. (1988), "Coordination of aerodynamic and phonatory processes in fluent speech utterances of stutterers", *J. of Speech and Hearing Research*, vol. 31, pp. 352-361.

[12] Postma, A. & Kolk, H. (1993), "The covert repair hypothesis: prearticulatory repair processes in normal and stuttered disfluencies", *J. of Speech and Hearing Dysfluencies*, vol. 36, pp. 472-487.

VARIABILITY IN THE ARTICULATORY KINEMATICS OF LIPS AND JAW IN REPEATED /pa/ AND /ba/ SEQUENCES IN ITALIAN STUTTERERS

Claudio Zmarich, Emanuela Magno-Caldognetto, Kyriaki Vagges
Centro di Studio per le Ricerche di Fonetica, C.N.R. -
Via Anghinoni, 10 35121 Padova, Italy

ABSTRACT

The kinematics of the closing gesture in bilabial stop consonants produced at two different rates by stutterers and controls was investigated. The results show that stutterers score lower than controls in displacement and velocity of movement and that some dynamic measures differentiate the speech production of stutterers from controls.

INTRODUCTION

Kinematic investigation is more useful than perceptive or acoustic analyses because it can detect the possible minimal anomalies in the fluent speech behaviour of stutterers, and access to levels of decreasing speech variability [1]. Bilabial stop consonants production in stutterers was investigated to verify the hypothesized different responsiveness of stutterers to voiced/voiceless contrasts and to changes in speaking rate [2].

PROCEDURE

Four stutterers (mean age: 25.25) and four normal subjects (mean age: 28.50) participated in the experiment. All the subjects had a negative history concerning neurological, speech, language and hearing problems, except for stuttering. Stuttering severity, as assessed by the Stuttering Severity Instrument [3], was mild for one subject and severe for three. One subject was never treated for stuttering, two subjects stopped the treatment 7 years before the date of the experiment and one subject 2 years before. All the subjects were instructed to repeat each of the /pa/ or /ba/ syllables in ten sequences and two rates, normal and maximal, with evenly stressed syllables, in random order. For each sequence, the acquisition time was set to 2 seconds. Thus, each subject produced 40 sequences, except for one stutterer that produced 20 sequences. The normal rate was the preferred rate for each subject, and the maximal rate was the fastest rate the subject was able to perform without altering the perceptual characteristics of the phones. Upper lip (UL), lower lip (LL) and jaw (J) movements were recorded and analysed with ELITE, a fully automatic, real-time system for 3D kinematic data acquisition which uses small, non obtrusive, passive markers of 2 mm in diameter attached onto the speaking subject's face. This system ensures high accuracy and minimum discomfort to the subject [4].

In this study, the movements of the markers placed on the central points of the UL, LL and J were analysed. The LL movement was then digitally subtracted from the J movement. Interlabial vertical distance was recorded as the distance between UL and LL, providing a measure of the combined movement (C). Relevant data were then selected from the general tables reporting all the movement and velocity peaks and the acoustic signal segmentations, and considered for statistical analysis.

As our purpose was to consider only perceptively fluent utterances, we paid particular attention to signs of disfluency or defective articulations. Only 9 gestures in different sequences produced by one stutterer were eliminated due to slurred speech. Gestures were eliminated when one or both of the following conditions occurred: irregular movement form and frequency for those movements having less than one millimetre of amplitude; presence of more than one peak in the velocity curve referring to a gesture. Moreover, the steady state portions of the movements, mainly corresponding to the open mouth position, were not measured. For the stutterer group, the percentage of eliminated cases was 23.28 out of a total of 4725 gestures (1575x3 articulators). For the control group, the percentage was 20.18 out of a total of 5310 gestures (1770x3 articulators). In order to assess the spatial and temporal characteristics of

UL, LL, and J during the opening gesture and the closing gesture, the following measurements were taken: a) duration of opening or closing gesture, measured as the time interval between onset of the movement and peak opening or closing position; b) time interval from onset of opening or closing movement to peak velocity; c) displacement, calculated as the distance between onset position and peak opening or closing position; d) peak velocity.For each of these measurements, the effects of the voiced/voiceless contrast and of the different speaking rates on the normal/pathological condition of the subjects were analysed. As to the latter variable, speaking rate was classified in two ways. The first was subject-dependent, i.e.both the normal and the fast rates were related to the subject's own speaking style. The second was a post-hoc rearrangement of the original rates. In fact, to provide an objective rate-dependent group comparison, reference to the duration of C was considered necessary. Four classes were created: 0.050-0.100 (very fast), 0.101-0.150 (fast), 0.151-0.200 (moderate), 0.201-0.250 (slow).Outliers were eliminated.

RESULTS

The mean number of gestures analysed for each stutterer was 123.2 for the preferred rate condition and 270.5 for the fast rate condition. For the controls there were 164.5 and 278.5, respectively. The mean duration of the C gesture provided a measure of the articulatory rate. At the preferred rate, the stutterers produced 5.29 gestures per sec. and the controls 5.55 gestures. At the fast rate, the stutterers produced 8.71 and the controls 9.00 gestures per sec. Only the data relative to the closing gestures are presented here, as this gesture appeared to differentiate the stutterers from the controls better than the opening gesture [5]. The data obtained with the subject-dependent rate will be presented first. Statistical analysis was applied involving a planned series of separate comparisons between stutterers and controls within each rate using the Mann-Whitney survey ($p = .005$). A non-parametric survey was chosen because of the non-normal distribution of the data (different variance, different number of cases and presence of ratio data).Tab.1 shows the median values and the significant comparisons between stutterers and controls for gestures displacement, velocity and duration. In addition to these direct measures, indirect measures are also presented: peak velocity / displacement ratio (a measure of the mass-normalized stiffness, cf. [6]); time from movement onset to peak velocity / total movement time x 100 (a measure of the simmetry of the velocity profile, cf. [7]), and parameter c (a metric of the velocity profile shape, cf. [6]). The formula of paramenter c is :(peak vel. / displacement) x movement duration. For these indirect measures the normative studies [6,7] established the following trade-off between rate and scores: both vel./displacement ratio and % of time to peak velocity vary positively with rate, while parameter c varies inversely. For the % of time to peak vel., our data show an interesting counterevidence to the norms, probably because we did not count the steady state portion of the gestures and the multiple-peak velocity gestures, which are much more frequent at slower rates.

The stutterers perform the gesture with less amplitude and velocity compared to the controls, while duration is less affected. The differences between groups for /p/ are more significant than /b/. The data of velocity/displacement ratio are greater for the controls than for the stutterers, because the velocity is proportionally higher in the controls. Considering the % of time to peak velocity, the general trend for stutterers is to have higher values than controls. Generally speaking, all these effects are more evident for J and less for UL (see Tab.1).In interpreting these data, however, we must take into account the intersubject variability, in part due to the task itself, as the subjects articulated according to personal feelings of comfortable and maximal rates. A way of minimizing these effects was to relate all the kinematic values of each subject to four classes, based on the duration of C.Unfortunately, as individual data were not equal in number across the four categories, statistical comparisons were precluded

Table 1.Comparisons of the kinematic measures of the movement across speaking rates. Measures included displacement (D: mm), duration (T:sec), peak velocity (V: mm/sec), peak vel./displ. ratio (R), parameter c. (P) and % of time from onset to peak velocity (%). Values were compared within each rate using the Mann-Whitney statistic (p=.005).

Stutterers vs Controls			/p/ closing						/b/ closing					
Art.	Rate		D	T	V	R	P	%	D	T	V	R	P	%
UL	Preferred	S	1.14	0.180	13.5	11.50	1.90	57.1	1.19	0.170	14.9	11.22	1.78	56.6
	Preferred	C	1.62	0.140	21.7	13.16	1.84	53.8	1.28	0.145	17.5	12.52	1.77	57.1
	*p=.005			*	*	*				*		*		
UL	Fast	S	2.17	0.120	29.2	13.9	1.67	53.8	1.45	0.110	20.7	16.2	1.70	53.8
	Fast	C	1.21	0.120	18.2	14.5	1.74	50.0	0.87	0.110	15.5	16.5	1.69	50.0
	*p=.005							*	*		*			
LL	Preferred	S	1.14	0.170	20.7	17.26	3.00	56.2	1.30	0.170	20.2	14.40	2.52	57.8
	Preferred	C	2.35	0.170	40.6	17.00	2.35	50.0	2.41	0.190	43.2	17.33	3.21	56.2
	*p=.005		*		*			*	*		*	*	*	
LL	Fast	S	0.59	0.100	15.0	22.2	2.20	50.0	0.81	0.100	22.35	21.57	2.26	50.0
	Fast	C	1.86	0.110	37.8	19.7	2.04	45.4	1.98	0.100	41.68	20.10	2.16	50.0
	*p=.005		*	*	*			*	*		*			
J	Preferred	S	6.60	0.170	73.4	10.77	1.91	55.2	7.27	0.180	69.8	14.40	1.88	57.5
	Preferred	C	7.84	0.180	111.8	11.59	2.19	52.9	8.24	0.190	109.9	17.33	2.22	56.2
	*p=.005		*		*	*	*	*	*		*	*	*	
J	Fast	S	5.02	0.100	86.1	16.17	1.62	45.8	5.25	0.100	88.7	15.62	1.62	50.00
	Fast	C	6.38	0.110	99.9	15.45	1.69	50.0	6.45	0.100	120.2	17.24	1.72	50.00
	*p=.005		*	*	*		*	*	*		*		*	

As an example, when the data for the J closing gestures are plotted on a three-dimensional space, with velocity and displacement related to different rates and the graphics of the subjects are paired according to comparability of number of occurrences for the different rates, a picture of extreme inter-subject variability appears. Apart from this variability, the values for the J movements in the production of /ba/ (b) are lightly higher than /pa/ (p) for most of the subjects. While the general trend is to reduce the amount of displacement and velocity along with the increase in speaking rate, the stutterers S2 and S3 are the only subjects that present some increase of displacement and velocity. To conclude, the dramatic contrast between S1 and C1 shows us that different speakers are able to perform the same phonetic gestures at comparable rates with an extremely divergent use of the articulators, and they do this without becoming disfluent.

REFERENCES

[1] Alfonso, P.J. (1991), "Implications of the concepts underlying task-dynamic modeling on kinematic studies of stuttering", in H.F.M. Peters, W. Hulstijn, C. W. Starkweather (Eds.), *Speech motor control and stuttering*, Amsterdam, Excerpta Medica, p. 79-100.

[2] Bloodstein, O. (1987), *A handbook on stuttering*, Chicago, National Easter Seal Society.

[3] Riley, G.D. (1972), "A stuttering instrument for children and adults", *Journal of Speech and Hearing Disorders*, 37, pp.314-322.

[4] Magno-Caldognetto E., Vagges K., Ferrigno G. & Zmarich C. (1993), "Articulatory dynamics of lips in italian /'VpV/ and /'VbV/ sequences", *Proceedings of Eurospeech '93*, vol.1, Berlin, pp. 409-412.

[5] Zmarich, C., Magno Caldognetto, E., Vagges, K., "Articulatory kinematics of lips and jaw in repeated /pa/ and /ba/ sequences in italian", to appear in *Proceedings of the first world congress on fluency disorders.*

[6] Ostry, D.J. & Munhall, K.G. (1985), "Control of rate and duration of speech movements", *Journal of the Acoustical Society of America*, 77, pp.640-648.

[7] Adams, S.G., Weismer, G., Kent, R.D. (1993), "Speaking rate and speech movement velocity profile", *Journal of Speech and Hearing Research*, 36, pp.41-54.

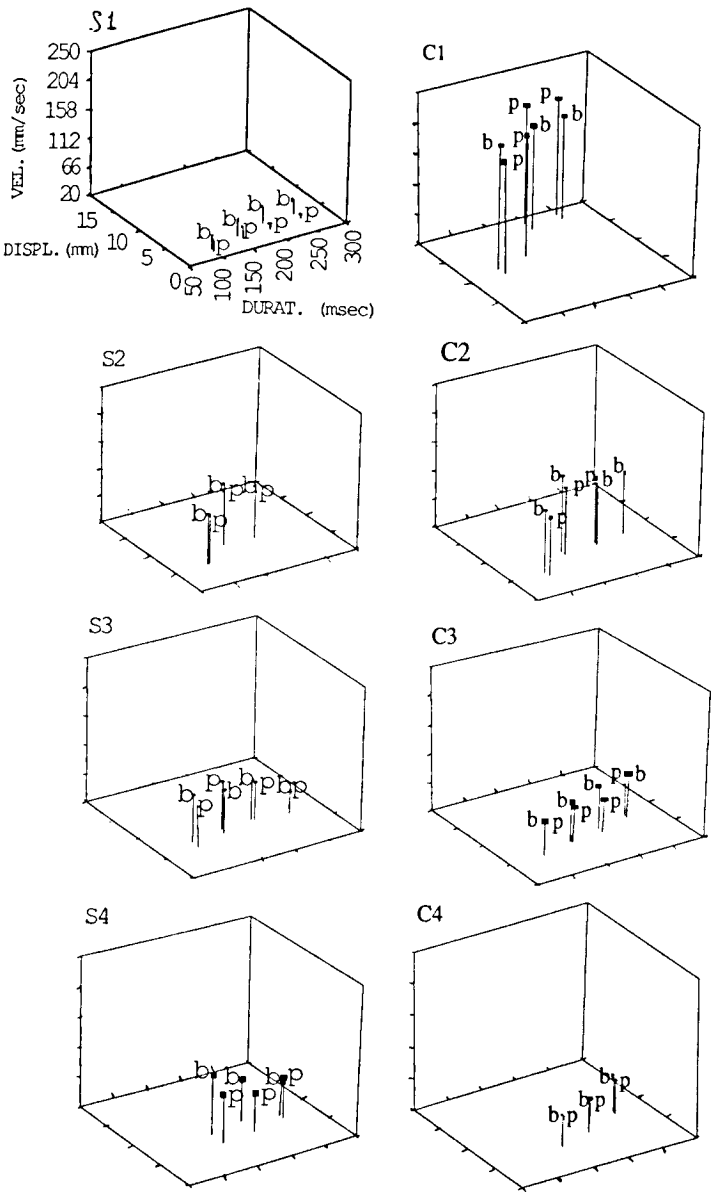


FIG.1. *Individual peak vel. and displacement means of the closing gestures for J as a function of four different rates for /p/ and /b/: 100:50-100; 150:101-150; 200:151-200; 250:201-250.Stutterers (S1, S2, S3, S4) and Controls (C1, C2, C3, C4) are paired according to comparability of number of occurrences along with the rates (S1-C1 produced most gestures for 100, and so on).*

THE PHONETIC GOALS OF THE NEW BAVARIAN ARCHIVE FOR SPEECH SIGNALS

Hans G. Tillmann, Christoph Draxler, Kurt Kotten, Florian Schiel
IPSK – Institut für Phonetik und Sprachliche Kommunikation, Munich, Germany
{tillmann,draxler,kotten,schiel}@sun1.phonetik.uni-muenchen.de
http://bas.phonetik.uni-muenchen.de/

ABSTRACT

The paper describes the phonetic motivation and orientation of the new publicly funded Bavarian Archive for Speech Signals (BAS).

The BAS collects, evaluates and makes accessible very large corpora of Spoken Language (SL-) corpora. These corpora will serve to develop a (more or less) Complete Phonetic Theory (CPT) of spoken German (CPT in the sense of [8]).

INTRODUCTION

Two general arguments of principal theoretical and practical interest played a major role in the decision to found a new archive for collecting spoken utterances of German, the Bavarian Archive for Speech Signals (BAS).

The first argument relates to the theoretically yet unresolved issue to what extent speakers (when producing utterances for conducting speech acts) follow the generative rules of a grammar or rather just produce copies of templates. In the latter case they have learnt to use a set of stereotypes and adopt them semantically (by selecting words from their mental lexicon) and modify them by transformation rules (such as putting in an appropiate pronominal).

How creative speakers are in making use of their language is still an open question which can only be answered by the investigation of very large corpora of empirically collected genuine speech utterances. Here, the theoretical issue of principal interest is whether the language model behind the data is best to be described by statistical or by logical (i.e. rule-based) methods.

The second argument concerns the necessity of collecting, evaluating, and making accessible very large corpora of naturally spoken utterances for the development of technical applications in the domain of Spoken Language Processing (SLP). Not only do the phonetic sciences assist and support the development of SLP technology. It is at least as important to apply SLP-methods to speech research in order to produce new phonetic knowledge.

MANAGEMENT OF VERY LARGE SL-CORPORA

SL-corpora consist of the digital speech signal and associated symbolic annotation and administration data, such as orthographic or phonemic representations of the utterances, technical specifications of the recording equipment, speaker information), or other related information.

At present, most SL-corpora are distributed on storage media, e.g. magnetic tapes or CD-ROM. Signal data is encoded in a variety of (possibly proprietary) signal file formats and symbolic data representations. All data is stored in file systems which are operating system dependent.

Clearly, this approach to SL-corpora structure and dissemination is limited:

- The size of today's SL-corpora exceeds by far the storage capacity of distribution media (TED: 7 CDs, PhonDat: 7 CDs, Verbmobil: 6 CDs and growing).
- The lack of a standard for the representation of symbolic data leads to incompatible annotation systems, making the re-use of corpora in new contexts impossible.

SL-corpus server

A new approach to making data available consists in providing accountable network access to an *SL-corpus server*. Clients of the SL-corpus server must register to be allowed access to the data. They can either download the parts of the corpus they are interested in, or access the server on-line.

This approach has several advantages:

- Only relevant subsets of the corpus need to be accessed.
- Fine-grain control of user access to data is possible.
- Updates of a corpus become immediately available.
- Clients are shielded from storage and implementation details of the corpus data.

Network requirements

On the technical side, the SL-corpus server requires access to high-speed networks. The ISDN bandwidth of 2 x 64 Kbit/s should be considered as the lower limit and used for downloading only (uncompressed downloading a 600 MB CD will take approx. 12 hours). High-speed networks (> 100 Mbit/s) will be necessary for on-line access. Such networks are currently being deployed in Western Europe and the United States.

Data modelling requirements

On the data modelling side, the SL-corpus server requires a standard for the symbolic data representations, e.g. the computer representation of individual languages (CRIL) guidelines agreed upon at the IPA Kiel 89 convention (representation of speech data on three symbolic levels: orthography, citation form, and phonetic transcription) [2]. Furthermore, mappings of alphabets and coding systems are needed to be able to integrate data from different sources.

A first such database is the current PhonDat-Verbmobil Database, which is implemented in a persistent Prolog environment [1].

BAS technology

Since high-speed networks are not yet available everywhere, the BAS will continue the traditional dissemination of SL-corpora on CD-ROM for at least the next two years.

Currently, all SL-corpora are stored on a large central archive storage at the Leibniz-Rechenzentrum (LRZ), to which the IPSK is connected via a 100 Mbit/s fiber optic link. This archive has a capacity of approx. 2 TB (Terabyte) and a 10 GB cache.

An experimental setup to access this archive is now installed. It uses the Andrew File System (a uniform file system over multiple machines) to provide access to the data, and it supports user services such as retrieving data into the cache at a pre-specified time so that it can be accessed quickly.

A case study of using a DBMS to make available the BAS corpora is scheduled to start in summer '95.

THE WORD AS A CENTRAL PHONETIC UNIT

A narrow phonetic transcription of a human utterance is of little use if we do not know which words of which language the speaker intended to express. Even if we consider the phonetic description of single speech sounds it is important to understand that the segmental components of speech utterances could attract the interest of speech research only after alphabetic elements had become available in the form of words pronounced in phonetic citation forms [7]. Therefore the definition of the phonetic goals of the BAS was based on the decision to

consider the word to be the central phonetic unit of speech research.

Single words are the first thing a new speaker of a language has to learn, and any fluent speaker of a language can easily select a word from a connected speech utterance and demonstrate it in isolation to himself and his audience. If articulated in a clear and careful pronunciation, we get citation forms which are the models for the lexical entries described in pronuciation dictionaries. Both the great stability and consistency of citation forms and the great phonetic variability of connected forms explain why the factual phonetic form of actually pronounced words remain so unconspicous to the untrained speaker and listener.

"Les modifications phonétiques du langage" [4] had to be discoverd by the first instrumental phoneticians one hundred years ago, and they cannot be ignored by today's linguists and speech researchers because they are the true source of all the difficulties in SLP.

It is our basic theoretical assumption that the factual phonetic form of any word is a computable function of a lexically given predicate that takes a segmental structure and a prosodic shape as (a contextually independent) input. It produces an output which is context sensitive because it has to take into account as two intervening variables the context of a prosodic phrase and a context of situation.

Citation forms are computed in a specific zero context with the prosodic shape of a one-word phrase (with a terminal or an enumerating F0-contour). At the same time they also specify the functional input for computing the phonetic word form of a given connected speech utterance. Thus it makes sense to take an abstractly defined canonic citation form and relate this to the actually given pronunciations in the data base. The annotation of BAS data therefore strictly adheres to the CRIL conventions.

A first example of how citation forms are systematically varied to be matched to a given speech signal using an HMM-based speech verification system is described in [5].

To determine the proper predicates and the algorithms for computing the sound streams of word sequences in connected speech utterances it is necessary to be able to relate the acoustic speech signal to the articulatory production. Many purely prosodic reflexes of reduced segmental structures can only be understood if we look at individual sound gestures and their systematic reduction to allegroforms (cf. the examples given in [7]).

Therefore, multi-sensor data are of great interest and should thus be incorporated in SLP-corpora (see the final chapter for details). Concerning the relation between speech production and the digital speech signal it makes no theoretical difference whether articulatory or acoustic representations are used because today we are in a position to compute the acoustic output from the articulatory geometry.

Only if we take the word as the basic phonetic unit of speech research will we be able to understand the information-bearing variation that determines the actual form of the lexically given parts of real speech. It is the final aim of our approach to develop a theory that explains the dynamic nature of word identity in agreement with concepts such as the syllabically organized „Ausprägungscode" proposed in [6] or of the „H and H theory" proposed by Lindblom [3].

BAS OVERVIEW

The BAS is a publicly funded institution formally associated to the Institute of Phonetics and Speech Communication of the University of Munich.

Personnel

The BAS members are Chr. Draxler who is responsible for network access and databases, K. Kotten for system administration, and F. Schiel for automatic evaluation and distribution of corpora.

Access

The BAS can be reached under the e-mail address: `bas@phonetik.uni-muenchen.de`

WWW access, including demonstrations of the corpora that are available, is possible under `http://www.phonetik.uni-muenchen.de/`

Services

The BAS provides the following services:

- Collection, evaluation, and dissemination of SL-corpora
- Customizing corpus subsets according to user specifications
- Development of corpora tools

Corpora

The BAS currently (April 95) offers the corpora listed in table 1:

Name	# Spk	# Utt	Charact.
SI1000	10	10.000	dictation
SI100	101	10.000	dictation
PhonDat I	201	16110	di-phone
PhonDat II	36	2007	train enquiry
VM 1.0.3	126	1840 turns	spontan. speech
VM 2.0	162	1538 turns	spontan. speech
TED 93	188		Eurospeech recordings, (including laryngogr)

Table 1: BAS corpora (April '95)

Corpora under development are

- VM 3.0, 4.0, and 5.0
- ERBA, a very large collection of train enquiries
- WD: phone-balanced read speech
- „Challenge Corpus", a collection of speech data that reflect problems in speech science and technology
- Polyphone-like telephone speech
- EMA, electro-magnetic articulography data of 3000 reproductions of the 15 German vowels in a defined CVC-context produced by 7 speakers in clear speech as well as in isolation

Information on these corpora can be obtained via e-mail or WWW.

REFERENCES

[1] Draxler, Chr. (1995): Introduction to the PhonDat-Verbmobil Database of Spoken German, PAP Conf. 95, Paris.

[2] IPA. (1989): The IPA Kiel Convention Workgroup 9 report: Computer coding of IPA symbols and computer representation of individual languages. *Journal of the International Phonetic Association* **19**, 81-82.

[3] Lindblom, B. (1990): Explaining phonetic variation: A sketch of the H and H theory. in: W. J. Hardcastle et al (eds): *Speech Production and Modelling*

[4] Rousselot, P.J. (1891): Les modifications phonétiques du langage, *Revue des patois gallo-romans* 4, 65-208.

[5] Schiel, F., Wesenick, M.-B. (1994): Applying Speech Verification to a Large Data Base of German to obtain a Statistical Survey about Rules of Pronunciation, *Proceedings of ICSLP 1994*, 279 - 282, Yokohama.

[6] Tillmann, H.-G. (1963), Das phonetische Silbenproblem. Phil. Diss, Bonn.

[7] Tillmann, H.-G. (1995), „Kleine und Große Phonetik", in press.

[8] Tillmann, H.-G. , Pompino-Marschall, B. (1993): Theoretical principles concerning segmentation, labelling strategies, and levels of categorical annotation for spoken language database systems. *EUROSPEECH 1993*, Berlin.

A COMPARISON OF GERMAN NAMES AND GERMAN WORDS

A. Mengel
Institute of Communications Research, Technische Universität, Berlin

ABSTRACT

German names are more difficult to read and write than German words. This paper presents evidence for this fact that allows for explanations of this behaviour by investigating the frequency, orthography, phonetic structure and interplay between written and spoken words. The consequences of the findings for future automated processing of names are recommended.

INTRODUCTION

It is commonly accepted that the relationship between the orthographic and phonetic structure of German names is more difficult to handle than that of German words. This has negative effects on speech synthesis and speech recognition systems on the one hand and on the use of German orthography and pronunciation by speakers on the other.

DATA

To substantiate and localise the reasons of the above experience, selected properties of four categories of proper names (christian names, surnames, street names, town names) and non-inflected nouns were compared. Non-inflected nouns were chosen, as they are that group of words which resemble names most in morphological, syntactical and semantical behaviour: They are not inflected, they can act as subjects and objects and they denote entities. Names are rarely inflected, thus, only non-inflected nouns were chosen. The names and their transcriptions were taken from the German part of a CD-ROM produced by LRE-Project of the European Community called ONOMASTICA [1]. The CD-ROM contains approximately 2,000,000 German proper names. Only data which had been checked by humans were chosen. The non-inflected nouns and their transcriptions were supplied by the CELEX lexical database [2]. Entries without frequency information were not taken into account.

The transcriptions of the data include information on the sounds, syllable boundaries, primary and secondary stress. The data from the CELEX were adapted to the transcription standard used in the ONOMASTICA data. Consonant and vowel clusters surrounding syllable boundaries were standardised across all kinds of entries.

Table 1 shows the categories chosen, the number of the entries taken, their cumulative frequency of occurrence, and the coverage of the selected data.

Table 1. Number, frequency and coverage of the selected data.

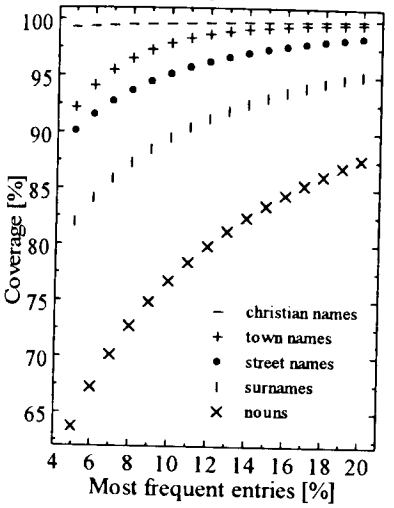
Cat.	n	Frequency	Cov. [%]
chri	10,778	33,040,984	97.62
surn	51,473	22,423,645	67.07
stre	73,605	34,182,103	63.82
town	25,892	40,570,131	99.34
noun	18,713	763,850	100.00

FREQUENCY OF OCCURRENCE

The number of different entries in the ONOMASTICA-corpus differs from category to category; also the number of items that are needed to reach a certain amount of coverage changes widely among types of entries. This can be seen best in figure 1. It shows how many of the most frequent entries of a name category are needed (x-axis) to cover a certain amount of entry frequency (y-axis). With 16% of the most frequent nouns, 85% of all occurrences of nouns are covered. The deviation of the graphs from an assumed straight line can be interpreted as a measure for the unevenness of distribution of entries in a given kind of name or the nouns and has the following implication: It is difficult to predict the occurrence of an entry in surnames or street names, but it will be easier in christian and town names because the likelihood for some entries is very high and is low for others. Thus, it seems that it is even more difficult to predict a certain noun than to predict a given surname. Yet, on the whole it will always be more difficult to predict a name than to predict a given word or noun: The usage of each individual noun in a given sentence is more restricted by syntactic and semantic constraints whereas there are less situations

in which a name is predictable by the context.

Figure 1. Percentage of most frequent entries and coverage of entries.



LETTER-TO-SOUND RULES

It is difficult to measure the difference of correspondences between orthography and pronunciation across different name and word categories. Even more difficult is it to measure the extent of such differences. With rule-based systems at hand one can only argue on the basis of individual string-categories or special letter-sequences and their corresponding pronunciation. Thus, only non-rule-based approaches can function as an instrument for this sort of investigation. Two self-learning methods were applied to make a measurement possible. The first is a neural-net-based system (BACK), the net type of which is a multi-layer perceptron. It is trained by backpropagation [3]. The second system is a self-learning system [4] based on a statistical approach (SELEGRAPH). Both systems were independently trained with the five different data sets. The resulting nets (BACK) and databases (SELEGRAPH) should then represent five different functions between orthography and transcription of the five categories.

To investigate the common assumption that the transcription of words and different categories of names need separate sets of letter-to-sound rules, a set of identical character sequences must be transcribed by the five different versions of the two self-learning algorithms.

Six-thousand strings with lengths of four to six characters were identified. They can be found as substrings in entries of all of the five data types. Also, entries of these lengths can be found in each of the data types.

These 6,000 strings were transcribed by the 5 different versions of each system. Markers for suprasegmental information were deleted from the transcriptions. Then, each transcription produced by one version of the net/database (e.g. the BACK-system version trained for christian names) was compared to those produced by versions for other categories of names or nouns (i.e. the BACK-system versions for surnames, street-names etc.).

Table 2 shows the percentage of the character sequences transcribed differently by a pair of differently trained systems for the neural net (BACK) and the statistical approach (SELEGRAPH). Figures of deviations are ordered in descending order.

Table 2: Percentage of differently transcribed strings.

BACK		SELEGRAPH	
diff. [%]	pairs	diff. [%]	pairs
60.55	chri-noun	40.23	chri-noun
57.45	chri-stre	34.87	stre-noun
50.78	chri-surn	34.78	surn-noun
50.52	chri-town	32.08	town-noun
44.97	stre-noun	31.60	chri-town
44.05	town-noun	31.11	chri-stre
41.58	surn-noun	28.85	chri-surn
37.53	surn-stre	16.43	surn-town
36.65	stre-town	16.15	stre-town
28.95	surn-town	14.65	surn-stre

Supposed, the percentage of differently transcribed entries is interpreted as measure for the difference between the LTS-correspondence of two separate entry types. It is then obvious that the difference of christian names and nouns is biggest (first row). However, the difference between surnames, street names and town names is smallest (last three rows). From the results, it is indeterminable whether the differences between christain names and the other name categories, or the difference between the nouns and the

other names is bigger. Hence, christian names and nouns seem to mark the edges of the range of correspondences between German orthography and pronunciation.

MINIMAL PAIRS

The concept of minimal pairs is used to evaluate the phonetic similarity of a group of entries. Minimal pairs are pairs of words of equal number of sounds that differ from each other in one sound only. In this investigation, diphthongs were treated as long vowels, affricates as two sounds, and the glottal stop as a consonant. Table 3 shows the percentage of entries that have minimal-pair partners.

Table 3. Percentage of entries that have minimal-pair partners.

Category	Entries with minimal-pair partners [%]
chri	59.69
sum	70.98
stre	46.32
town	44.04
noun	19.58

All of the names have more minimal - pair partners than the nouns have. Thus, in a speech recognition system the recognition performance for nouns would be better than it would be for any name. It would be worst for surnames and christian names.

HOMONYMY

In order to have another look on the orthography of names and nouns, the number of different orthographic strings in the corpus under investigation is measured against the number of different phonetic strings. The ratio of different phonetic strings and different orthographic strings (p/o) is calculated.

Figure 4. Ratio of different phonetic and orthographic strings.

Category	p/o [%]
chri	87.86
sum	85.83
stre	92.19
town	97.61
noun	99.00

These results rather provide information on the relation of orthographic and phonetic structures of the entry types while minimal pairs only express something about the phonetic aspects of entries: The more homophones there are, i.e. orthographically different entries with the same pronunciation, the more difficult will it be to determine the correct orthography of a transcription or pronunciation. Hence, it is more difficult to find the correct entry. Again, this has severe impact on speech recognition.

CONCLUSION

Four aspects of differences between four types of German names and non-inflected nouns have been addressed: the frequency of occurrence, LTS-correspondences, the number of minimal pairs and homonymy. All of them show that especially surnames deserve particular attention and require more effort for processing, be it human or automatic. Thus, for the implementation of future applications that include the use of personal names, more refined methods must be developed to cope with the state-of-the-art performance achieved for words.

REFERENCES

[1] ONOMASTICA (1995): *Transcription database for proper names of 11 European languages,* Edinburgh: CCIR, University of Edinburgh.

[2] Baayen, R.H.; Piepenbrock, R. & van Rijn, H. (1993), *The CELEX Lexical Database (CD-ROM),* University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

[3] Rosenke, K. (1994), 'Einsatz von neuronalen Netzen zur Transkription von orthographischem Text in Lautschrift', *Konferenz 'Elektronische Sprachsignalverarbeitung'*, Technische Universität Berlin, Institut für Fernmeldetechnik, pp. 460-467.

[4] Andersen, O. & Dalsgaard, P. (1994) 'A Self Learning Approach to Transcription of Danish Proper Names', *Proceedings of the International Conference on Spoken Language Processing*, Yokohama, pp 1627-1630.

THE DEVELOPMENT OF PHONOLOGICAL BIAS: PERCEPTION AND PRODUCTION OF SWEDISH VOWELS AND TONES BY ENGLISH SPEAKERS

Denis Burnham and Camilla Torstensson
School of Psychology, University of NSW, Sydney, Australia

ABSTRACT

English-speaking adults and children were tested for perception and production of Swedish vowels and tones. Six-year-olds showed greater reliance on vowel than tone distinctions, and more perceptual flexibility with vowels than older subjects, showing the emergence of a perceptual bias between 6 and 10 years. However, bias in speech production appeared later, around 14 years.

INTRODUCTION

Infants can discriminate many consonant contrasts, even those phonologically irrelevant in the ambient language [1]. A major part of speech perception development thus involves the *loss* of perceptual ability for irrelevant speech sounds. There appears to be two developmental periods in which this loss occurs. In early infancy, Werker [2] established that 7-month-old English language environment (ELE) infants perceive various Hindi and Salish Indian consonant contrasts but this ability deteriorates between 7 and 11 months. The second period of loss seems related to the onset of reading: Burnham [1] found inferior performance on non-native consonant contrasts by 6-year-olds compared with younger and older children, and a positive relationship between reading and phonological bias - children who were good readers were also better at perceiving native than non-native contrasts.

More recently work has been conducted on developmental processes in vowel and tone perception. This is important in the context of second language learning because it seems that foreign accents are carried mostly on vowels and tones. Kuhl [3] found that by 6 months infants have established prototypes for native vowels and, in what is called the "magnet effect", that nearby non-native vowels are absorbed into these prototypes. Kuhl found a perceptual drift for American infants towards the English vowel /i/, and for Swedish infants towards the Swedish vowel /y/. With regard to tones, there is some evidence that tonal distinctions are more functionally salient for infants than consonantal distinctions [4]. However, in children from non-tonal language environments tone perception appears to be relatively depressed by 6 years: Burnham and Francis [5] found that ELE 6-year-olds were better at discriminating a Thai non-native consonantal contrast than Thai tonal contrasts, while for adults the opposite was true. They suggest that, due to phonological bias, non-tonal language 6-year-olds have difficulty perceiving tones, despite their high acoustic salience.

Here we investigate the development of Swedish vowel and tone perception and production by English-speaking children and adults. Swedish was chosen because it has a tonal distinction and many vowels not found in English.

SUBJECTS

A total of 72 subjects were tested, 18 at each of four ages, 6 years, 10 years, 14 years, and adults. All were English speakers and none had experience with a tonal or a Scandinavian language. All 14-year-olds had reached puberty but no 10-year-olds had. All participated in both a perceptual discrimination and then a perceptual identification task, and 12 randomly-selected subjects at each age participated in a production task.

DISCRIMINATION EXPERIMENT

Subjects were tested by laptop computer, which stored and presented sounds, on an AX discrimination task and had to respond "same" or "different" by pressing one of two keys mounted on a response box providing digital I/O to the computer. Each trial was initiated by pressing a "ready" key, after which the two sounds were presented separated by either 500 msec or 1500 msec. (Results were later pooled as analysis showed no difference between these intervals.)

All stimulus items were natural Swedish productions carried on the nonsense word [mɛb*n]. Three levels of vowel contrast difficulty were tested (near, medium, and far in terms of distinctive features). In each age group two sub-groups (n=9) were tested, one in which both members of the pairs of vowels were phonologically irrelevant to English speakers (Swedish-Swedish (SS) sub-groups), and one in which one member of each pair was irrelevant and one was the same as an English vowel (Swedish-English (SE) sub-groups). (Here slight phonetic differences between the two languages were ignored.) For the SS sub-groups, the contrasts were [y] vs [ɵ], [y] vs [ø], and [y] vs [o]; in the SE subgroups [y] vs [i], [y] vs [e], and [y] vs [ɑ]. In addition, the Swedish tone contrast, [mɛbýn] vs [mɛbỳn] was tested in both sub-groups. Two blocks of 16 trials (4 of each of the 4 contrast types) were presented. Three exemplars of each sound were available on disk and the program selected from these at random to minimise the effect of acoustic cues and maximise the salience of phonetic cues. The dependent variable was a discrimination index (DI) - the number of correct responses on different trials (hits) minus the number of incorrect responses on same trials (false positives) over the number of trials of each contrast type.

It was expected that 6-year-olds should discriminate vowels better than tones, while the reverse should be true for adults [6]; and that phonological bias should increase with age [1].

Mean DIs for SS vowels, SE vowels, and tones across ages are shown in Figure 1. An age x vowel group x (contrast type) analysis of variance (ANOVA) revealed that all subjects discriminated vowels better than tones and SS better than SE vowel contrasts. Inferior performance on SE vowels indicates that a magnet-type effect was occurring: the S vowel in SE pairs was assimilated into the nearby E vowel prototype, while for the SS vowel pairs, the unfamiliar vowels remained more distinct perceptually. Post-pubescent subjects discriminated all contrasts better than did pre-pubescent subjects as did 10-year-olds over 6-year-olds. Such general effects can be understood in terms of subjects' improving ability to attend and fulfil the requirements of the task. So it is *differential* changes over age which are most important to note. Of specific interest to the hypotheses, there were significant effects of vowels/tones x pre/post pubescence, $F(1,64) = 25.34$, vowels/tones x 6/10 years, $F(1,64) = 9.64$, and of vowels/tones x SS/SE x 6/10 years, $F(1,64) = 4.11$. These results show that there was greater improvement for tones than vowels between pre- and post-pubescence, and even earlier between 6 and 10 years. In addition, as there was greater improvement for SS than SE vowels between 6 and 10 years, it seems that 6-year-olds are showing less of a magnet effect and thus less phonological bias. This can be seen better in Figure 2, in which SS minus SE scores are shown

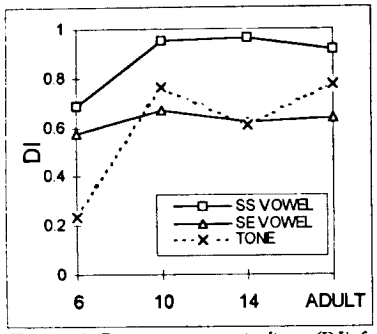


Figure 1. Discrimination indices (DI) for tones and vowels.

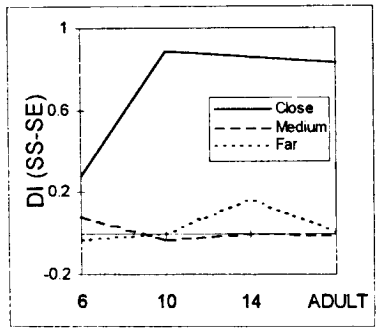


Figure 2. Magnet effect, DI (SS-SE) for close, medium, and far vowels.

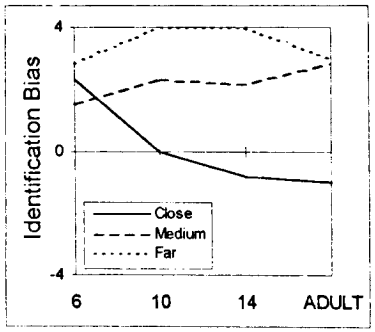


Figure 3. Identification bias (positive = vowel bias, negative = tone bias) for close, medium, and far vowels.

for the three vowel distances across ages. The ANOVA revealed interactions of SS/SE vowels and 6/10 years with the linear effect of vowel distance, F (1,64) = 8.91, and with the quadratic effect of vowel distance, F (1,64) = 12.52. When vowels are far or a medium distance apart, there is no superiority of SS over SE. However when vowels are close there is a large magnet effect for 10-year-olds and older subjects, but not for 6-year-olds. Thus, 6-year-olds show less phonological bias than older subjects for vowels (Figure 1), despite the fact that they appear to be biased against the use of tonal contrasts in a linguistic context (Figure 2).

IDENTIFICATION EXPERIMENT

If 6-year-olds are unable to use tone to distinguish lexical items, then this should show up in a task in which vowel and tone distinctions are functionally relevant. The same apparatus was used as in the discrimination experiment. However, here just a single sound was presented on each trial. In training trials subjects were presented with one of two sounds, eg., [mɛbýn] or [mɛbìn], which differed both in vowel and tone, and were required to press one of two buttons. Once they reliably identified these to criterion, 8 test trials were presented, 2 of each of the following: [mɛbýn], [mɛbìn] (the original training stimuli), and [mɛbỳn], and [mɛbín]. The latter two were designed to test whether the vowel cue or the tone cue was more salient for subjects. This training-test sequence was repeated twice so that subjects received a total of 8 novel test stimuli. As in discrimination, three versions of the task were employed for close, medium, and far vowels.

It was expected that subjects should base their identifications on vowels when the distance is great, but on tones when the vowel distance is reduced.

An age x SS/SE x vowel distance ANOVA revealed a significant linear trend over vowel distance, F (1,48) = 27.20, a significant linear x SS/SE effect, F (1,48) = 17.41, and a close to significant linear trend x 6- vs 10-year-olds, F (1,48) = 3.75. As can be seen in Figure 3, there is a definite vowel bias for far and medium vowels. For close vowels the 6-year-olds maintain their vowel bias, even though older subjects now rely more on tones. Thus despite a difficult vowel discrimination task, 6-year-olds are unable to use the presumably more salient tonal difference, due to their difficulty in attending to tonal distinctions in a linguistic context.

PRODUCTION

For production the subjects' task was to repeat various words modelled by a native Swedish speaker. The Swedish-only vowels [y], [ɵ], [ø], [o], and Swedish/English vowels [a], [e], [i] were presented in a [hVd] context. For tones, 'anden', 'biten', and 'tomten' were pronounced with either the single tone (falling on second syllable) (English = 'duck', 'the bit', and 'building site'), or with the double tone (rise on first and rise-fall on second syllable) (English = 'spirit', 'bitten', and 'santa claus'). The single tone was taken to be native in the sense that it uses a tone sequence familiar to English speakers and the double tone to be non-native. Two native Swedish speakers scored whether the vowels were correct and which of the tone words the subjects said. Subjects were better at native than non-native sounds and better at tones than vowels. Preliminary analyses show that the curves for native and non-native vowels are relatively parallel and flat across age, while for tones there is pre- to post-puberty improvement on the native tone and a reduction for non-native tones. The latter is consistent with the notion that in the perception tasks adults' superior performance with tones is due to the relatively high acoustic salience of tone differences compared with spectral qualities of vowels, rather than to any linguistic salience of tones. The results also provide some support for the notion that the ability to produce non-native speech sounds deteriorates after puberty

CONCLUSIONS

For phonologically-irrelevant vowels perceptual flexibility decreases markedly between 6 and 10 years. However, 6-year-olds are much *less* flexible with tones than are their older counterparts. Paradoxically 6- and 10-year-olds show equivalent ability in producing native and non-native tones, while 14-year-olds' and adults' show superior ability with native tones. Thus there seems to be little correspondence between English speakers' perception and production of Swedish vowels and tones.

REFERENCES

[1] Burnham, D. (1986). Developmental loss of speech perception: Exposure to and experience with a first language. *Appl. Psycholing.*, *7*, 206-240.

[2] Werker, J.F. (1991) The ontogeny of speech perception. In I.G. Mattingly & M. Studdert-Kennedy (Eds) *Modularity and the motor theory of speech perception.* Hillsdale, N.J., Erlbaum.

[3] Kuhl, P., Williams, K., Lacerda, F., Stevens,K. & Lindblom, B. (1992) Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606-608.

[4] Li, C., & Thompson, S. (1977). The acquisition of tone in Mandarin-speaking children. *J. Ch. Lang.*, *4*, 185-199.

[5] Burnham, D. & Francis, E. (1995) The role of linguistic experience in the perception of Thai tones. In T. Thongkum (Ed) *SouthEast Asian Linguistic Studies in Honour of Vichin Panupong* Bangkok: Chulalongkorn University Press.

ACKNOWLEDGMENTS

We acknowledge the multifaceted assistance of Ms Elizabeth Francis.

ACQUISITION OF THE ENGLISH /r/-/l/ CONTRAST BY JAPANESE SPEAKERS: EFFECTS OF TRAINING IN PERCEPTION ON PRODUCTION

Ann R. Bradlow and David B. Pisoni
Speech Research Laboratory, Indiana University, Bloomington, Indiana, U.S.A.

Reiko A. Yamada and Yoh'ichi Tohkura
ATR Human Information Processing Research Laboratories, Kyoto, Japan

ABSTRACT

This study investigates the effects of training in /r/-/l/ perceptual identification on /r/-/l/ production by adult Japanese speakers. Subjects were recorded producing English words that contrast /r/ and /l/ before and after participating in an extended period of /r/-/l/ identification training. Improvement in the Japanese trainees' /r/-/l/ productions as a consequence of training in perception was evaluated by a direct comparison of the pretest and post-test productions by a group of native American English listeners. The results showed significant perceptual learning for all subjects as a consequence of the training program. More importantly, this perceptual learning transferred to the production domain, implying a close link between perception and production.

INTRODUCTION

It is well known that the English /r/-/l/ contrast presents difficulties in both perception and production for Japanese speakers, for whom the contrast is neutralized [1]. Furthermore, this contrast has proved difficult to acquire in adult second language learners [2]. This finding has led to the claim that certain non-native phonetic contrasts may, in fact, be nearly impossible for adults to acquire [2]. However, this claim has recently been challenged by a new approach to perceptual learning [3]. Traditional methods of training non-native phonetic contrasts were guided by an attempt to draw the trainees' attention to the individual acoustic features that differentiate prototypical versions of the members of the target phonetic contrast. This approach has not been successful in promoting the acquisition of robust, English-like /r/ and /l/ categories by Japanese speakers [2]. In contrast, a new orientation that was designed to expose the trainees to a wide range of exemplars of the target categories, has proven very successful [3]. The guiding principle behind this approach is that, in order to develop robust and linguistically meaningful phonetic categories, trainees must be exposed to exemplars that incorporate the variability that characterizes the target category. The success of this "high variability" training procedure suggests that adults are indeed capable of learning to perceive new, difficult phonetic contrasts [3].

The present study builds on this previous finding by investigating the effect of perceptual learning on /r/-/l/ production . Since the perceptual training program involves no production training whatsoever, transfer of the perceptual learning to the production domain would provide new evidence for a close link between perception and production, and would therefore be of both practical and theoretical interest.

METHOD

Perception Training

All perception training and testing was done at ATR Human Information Processing Research Laboratories. Eleven monolingual Japanese adults were trained over a period of 45 sessions using the "high-variability" training program. The stimuli consisted of English /r/-/l/ minimal pairs produced by five native English speakers. These minimal pairs included words with the target phoneme in multiple phonetic environments. Thus, the training stimuli incorporated a wide range of category variability due to cross-speaker differences, as well as differences in phonetic context. The procedure used for the training sessions was a two-alternative forced choice identification task, in which the trainees heard a stimulus and identified it from an /r/-/l/ minimal pair. (For example, trainees heard "brush" and identified it from the pair, "brush-blush.").

In order to assess the trainees' improvement in identifying English /r/-/l/ minimal pairs, they performed a pre- and post-test at the start and end of the training period, respectively. In addition, the trainees performed two tests of generalization at the post-test phase of the experiment. These tests were designed to assess the extent to which the trainees could generalize the newly acquired /r/ and /l/ categories to stimuli they had not previously been exposed to. The first test of generalization presented new words produced by one of the talkers who produced the stimuli in the training set. The second test of generalization presented new words by a new talker.

Production Pre- and Post-test

In addition to the perception pre- and post-test, the trainees performed a production pre- and post-test. In this test, the trainees were recorded reading a list of English /r/-/l/ minimal pairs. The individual words were presented in random order and the subjects were given both visual prompts (standard English orthography) as well as an auditory model (a male, General American English speaker's production of the target word). The auditory model was provided in order to assist the Japanese trainees with the pronunciation of the rest of the word besides the /r/ or /l/. These trainee recordings were made in an anechoic chamber at ATR Human Information Processing Research Laboratories, and were digitized at a sampling rate of 22.05 KHz with 16 bit resolution.

Subjects

Eleven monolingual Japanese speakers (six males, and five females) served as subjects in the perceptual training program. A comparable group of eight Japanese speakers served as control subjects. None of the subjects had received any special English conversation training, although (as is typical in Japan) all had studied English since Junior High School (age 12 years). All subjects were recruited from Doshisha University, Kyoto prefecture, Japan.

The control subjects performed the perception pre- and post-tests, as well as the two tests of generalization; however, these subjects did not go through the training program. In addition to the perception tests, these control subjects also performed the production pre- and post-tests. The time lag between the control subjects' pre- and post-tests was identical to the time of the training program for the experimental subjects.

Evaluation of Trainee Productions

The Japanese trainees' pre- and post-test utterances were transferred to the Speech Research Laboratory at Indiana University, where they were converted to 12 bit resolution for presentation to American English (AE) listeners using a PDP-11 computer. The AE listeners performed a direct comparison task on pairs of pre- and post-test tokens. In this task, the AE listeners saw the target word in standard English orthography on a CRT monitor and then heard two versions (a pre-test and a post-test version) of the target word by a single Japanese trainee. The AE listeners responded by selecting the version that sounded "better," or "more precisely articulated" on a seven-point rating scale. On this scale a response of "1" indicated that the first version was "much better" than the second version; a "7" indicated that the second version was "much better" than the first version, and a "4" indicated that there was no difference between the two versions. In the presentation of these stimuli, each pretest - post-test pair was presented twice: once in each of the two possible orders (pretest then post-test, and vice versa).

Each Japanese trainee's pre- and post-test productions were compared by a separate group of ten AE listeners, for a total of 110 AE listeners. Control subjects' pre- and post-test productions were evaluated by an additional 80 AE listeners (ten for each of the eight control subjects). These AE subjects were all students at Indiana University and received course credit for their participation in this experiment. All subjects reported no history of speech or hearing impairment.

RESULTS

Perceptual Learning

Figure 1 shows the Japanese trainees percent correct identification for the pretest, post-test, and the two tests of generalization. The left panel shows the

results for the trained group, and the right panel shows the results for the control group. For the trained group, there was a significant improvement in accuracy from pretest to post-test ($t(10)=-7.38$ $p<.001$ by a 2-tail paired t-test), and this level of performance was maintained in the two tests of generalization. In contrast, the control group showed no difference from pretest to post-test ($t(7)=2.185$, $p=.065$ by a 2-tail paired t-test).

These data replicate the results of previous /r/-/l/ perception training studies using the "high-variability" training procedure [3]. This pattern of results indicates that the trainees did indeed show significant perceptual learning as reflected in the significant changes in performance for the experimental group. In the present study, we were also interested in investigating how this perceptual learning affected the subjects' ability to produce more native-sounding words that contrast in /r/ and /l/.

Transfer of Perceptual Learning to Production

Figure 2 shows the distribution of responses for the AE listeners' comparisons of the Japanese trainees' (left panel) and control subjects' (right panel) pre-and post-test productions. This figure shows the proportion of trials for which the AE listeners judged there to be no difference between the pre- and post-test productions (post=pre), for which they judged the pretest version better than the post-test version (post<pre), and vice versa (post>pre).

For both the trained and control subjects, there was a relatively small proportion of trials that received the post=pre response. This proportion was higher for the control subjects than for the trained subjects. More importantly, a far greater percentage of the trained subjects' productions received a post>pre rating than the reverse rating. This is seen in Figure 2 by the significant difference between the frequency of post<pre and post>pre responses for the trained subjects' productions ($t(10)=-3.018$, $p=.013$ by a 2-tail paired t-test). In contrast, for the control subjects, there is no difference in frequency of the post<pre and post>pre responses ($t(7)=-.625$, $p=.552$ by a 2-tail paired t-test). In other words, the AE listeners showed a preference for the trained subjects' post-test productions over their pretest productions; whereas, the AE listeners showed no such preference for the control subjects' post-test productions.

Thus, the post-test /r/-/l/ utterances of the Japanese subjects who went through the training program showed significant improvement over the corresponding pretest productions. This result demonstrates a transfer to the production domain of the perceptual knowledge that was acquired during the training program.

DISCUSSION

The results of this study show that, even with no explicit production training, the perceptual learning that resulted from the /r/-/l/ identification training transferred to the production of /r/ and /l/. From a practical point of view, this finding suggests that the acquisition of new phonetic contrasts in production, as well as in perception, can be facilitated by extensive training in perception alone. From a theoretical point of view, these data indicate a close perception-production link, to the extent that learning in the one domain transfers to changes in the other domain.

ACKNOWLEDGMENTS

We are grateful to Takahiro Adachi for technical support, and to Fernando Vanegas for subject running. This work was supported by NIDCD Training Grant DC-00012 and by NIDCD Research Grant DC-00111 to Indiana University.

REFERENCES

[1] Goto, H. (1971), "Auditory perception by normal Japanese adults of the sounds 'l' and 'r'," *Neuropsychologia* vol.9, pp. 317-323.

[2] Strange, W., & Dittman, S. (1984), "Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English," *Perception and Psychophysics* vol. 36, pp. 131-145.

[3] Lively, S., Pisoni, D., Yamada, R., Tohkura, Y., & Yamada, T. (1994), "Training Japanese listeners to identify English /r/ and /l/: III. Long-term retention of new phonetic categories," *J. Acoust. Soc. Am.* vol. 89, pp. 874-886.

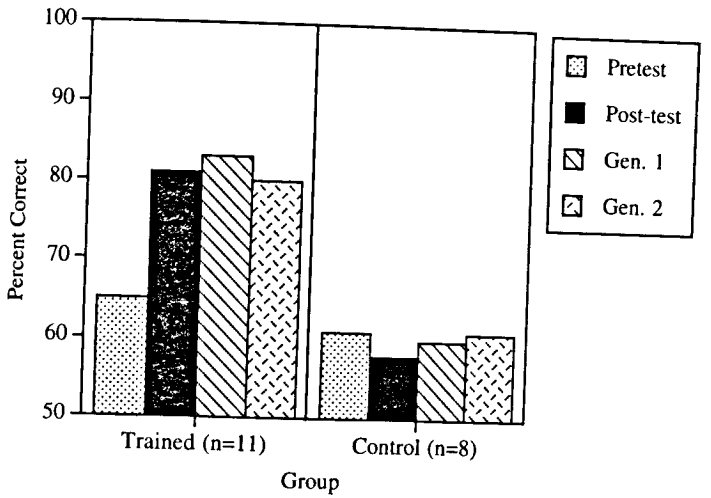


Figure 1. Japanese trained and control subjects' performance on the perceptual identification pretest, post-test and two tests of generalization.

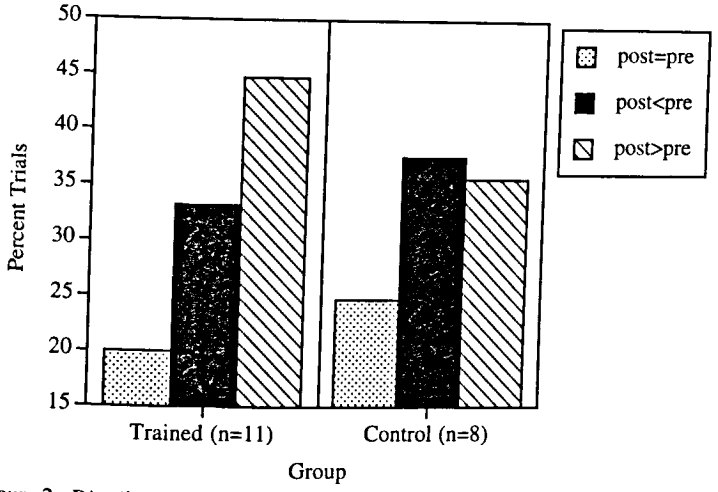


Figure 2. Distribution of AE listener comparison responses for the pre- and post-test productions by trained and control Japanese subjects.

PERCEPTION OF TENSE-LAX VOWELS AND FORTIS-LENIS CONSONANTS BY RUSSIAN LEARNERS OF ENGLISH

A.Yu. Panasyuk, I.V. Panasyuk, A.L. Gorlovsky, O.V. Anfimova
St.Petersburg State University, Philological Faculty, Laboratory of New Teaching Technologies

ABSTRACT

A perceptual experiment was conducted in a group of 76 Russian students of English in order to test their ability to identify tense-lax vowels and fortis-lenis stops in final position in monosyllabic words. The data obtained indicate that the listeners responded mainly to the durational characteristics of vowels and were less sensitive to their quality. In erroneous perception the subjects tended to follow the pattern of their mother tongue.

INTRODUCTION

The present study is part of a longitudinal project whose ultimate goal is to investigate perceptual ability in Russian University students of English at the intermediate level and to work out a diagnostic test enabling the teachers of English as a foreign language to quantitatively assess students' level of phonetic proficiency and administer remedial sets of auditory and oral drills.

It has been observed that two distinctive features of English phonemic system present a major difficulty to foreign learners of English (Russians included), namely, the distinction between the so-called "long" and "short", or "tense" and "lax", vowels and "voiced" and "voiceless", or "lenis" and "fortis", stop consonants [2, 4]. This difficulty may be accounted for both by the dramatic differences between the phonemic systems of Russian and English and by an intrinsic complexity of phonetic realization of these phonemic contrasts.

The discrepancy existing between Russian and English sound systems is clearly seen in CVC words. The English language system permits 4 word types differing in phonemic lengh of the vowel and presence or absence of voice of the final consonant, e.g. "bead", "beat", "bid" and "bit". In Russian, where the opposition of length is absent and that of voicedness/voicelessness is neutralized in word-final position, only one word /b'it/ is permitted, which makes the differentiation of the English contrasts for the Russian speakers quite a hard task. On the other hand, it is well known that the length of a vowel may vary considerably depending on the presence or absence of voice in the following consonant. Thus, vowels tend to be longer preceding voiced as compared to voiceless final tops [1, 3, 5]. As a result, the shortened /i:/ in beat is quite likely to be shorter than the long allophone of the "short" /i/ in "bid". The distinction between the phonologically "long" and "short" vowels is preserved by a difference in vowels quality rather than vowel duration.

While teaching English phonetics, we make our students aware of the fact that it is more convenient, for practical purposes, to use the terms "tense"/"lax" vowels and "fortis"/"lenis" stops since phonetic duration serves mainly as a means of differentiating final consonants.

MATERIAL AND PROCEDURE

The stimuli in this study were 48 monosyllabic words of the CVC structure which contained 3 vocalic contrasts: /i: - i/, /ɑ: - ʌ/ and /ɔ: - ɔ/ before t/d. The list of the words is given in Table 1. We assigned each word type a positional number, i.e. bead — Position 1, beat — Position 2, bid — Position 3 and bit — Position 4.

One can see that the vast majority of the test words are high frequency words.

However, we had to include some rare words, e.g. "fid" and some proper names, such as "Sid" and "Hudd". In one case we had to use an invented "name", "Stutt", because it was impossible to find a closely matched minimal pair to complete the set of the /ɑ: - ʌ/ contrasting words.

Table 1. The stimuli presented to Russian speakers of English for identification.

Vowels	Test words Position			
	1	2	3	4
i: - i	bead	beat	bid	bit
	feed	feet	fid	fit
	greed	greet	grid	grit
	seed	seat	Sid	sit
ɑ: - ʌ	card	cart	cud	cut
	bard	Bart	bud	but
	hard	heart	Hudd	hut
	starred	start	stud	Stutt
ɔ: - ɔ	pored	port	pod	pot
	cord	caught	cod	cot
	shored	short	shod	shot
	roared	wrought	rod	rot

The test words were read by a native speaker of English (a young man from Britain with a standard pronunciation) twice: first, in the order in which the words are presented in Table 1 (from position 1 to Position 4) and second, in a random order. In the latter case, each word was preceded by its number for the convenience of the listeners. Each word was pronounced only once. The speaker read the stimuli in a natural manner without exaggerating the production of the sounds to help the listeners, by lengthening the sound or using very explicit careful articulation. The interval between the words was not strictly defined but it was approximately the same throughout the list (2-3 sec.).

The material was recorded in a soundproof chamber and presented to the listeners through the headphones in the PRISMA AUDITEK language laboratory of the philological faculty of St.Petersburg State University.

The listeners were 76 students of the English department who had completed their second year of studies. All of them had had a two-year course in practical phonetics (2 academic hours a week). Their knowledge of English may be said to vary from lower to upper intermediate.

The perceptual tests were conducted in groups of 10-12 students. Each participant was provided with an answer sheet which contained the list of words ordered as in Table 1 for the training session and the answer sheet proper where each line contained the orally presented word and its three minimal pairs. The listener was to underline or encircle the word he or she thought was pronounced by the speaker.

RESULTS

Of the 76 subjects, 3 only (4%) fulfilled the task without any mistakes. All of them were very good students of English (upper intermediate or even advanced). 75% of the subjects performed very well, having yielded more 75% correct answers. The lowest percentage of correct answers was found to be 40% in 3% of the subjects.

As expected, the stimuli differed widely in the number of correct answers obtained. Table 2 gives the number of correct identifications in per cent for each position.

Table 2. Mean number of correct identifications for words with different vowels (%).

	Position number			
Vowels	1	2	3	4
i: / i	88	76	83	79
ɑ: / ʌ	91	83	68	83
ɔ: / ɔ	82	84	74	73

In Table 2 we can see that the words in Position 1 have the highest mean values of correct identifications. The highest

percentage was obtained for the word "bead" — 99% correct answers. The worst identified word was "Hudd"— Position 3 (40%).

The next step was to analyse confusions between the words. Confusion matrices were built for the words of the same vowel type. Table 3 shows substitutions of the presented words containing the vowels /i: - i/.

Table 3. Substitution matrix (in %) for the words with the /i:/ - /i/ vowels.

	words perceived				
word presented		1	2	3	4
	1	–	82	16	3
	2	15	–	37	49
	3	54	25	–	21
	4	8	41	51	–

Table 4. Substitution matrix (in %) for the words with the /ɑ:/ - /ʌ/ vowels.

	words perceived				
word presented		1	2	3	4
	1	–	77	14	9
	2	31	–	17	52
	3	36	21	–	43
	4	7	21	46	–

Table 5. Substitution matrix (in %) for the words with the /ɔ:/ - /ɔ/ vowels.

	words perceived				
word presented		1	2	3	4
	1	–	54	36	10
	2	27	–	19	55
	3	39	12	–	50
	4	5	69	27	–

Inspection of the matrices in Tables 3-5 indicates that the most pronounced tendency is for Position 1 to be substituted by Position 2. This tendency is a little weaker with the /ɔ: - ɔ/ vowels compared with the two other vocalic contexts. This substitution type involves one phonemic feature — devoicing of the final /d/. Interestingly, substitutions of the reverse direction occur much less frequently.

Another marked tendency is for words in Position 2 to change into words in Position 4, which means substitution of lax vowel for a tense one. In the case of the /ɔ: - ɔ/ vowels the reverse substitution prevails.

Position 3 words demonstrate a uniform substitution patterns — they are perceived either as Position 1 or Position 4 words.

Position 4 words are perceived either as Position 3 or Position 2 words.

All these substitutions involve 1 distinctive feature. Of the two possible two-feature substitutions one is clearly marked, namely, Position 2 perceived as Position 3. In terms of distinctive features it means (tense+fortis) → (lax+lenis) and vice versa, that is, (lax+lenis) → (tense+fortis), with a slight predominance of the former type of substitution.

On the contrary, the other two-feature substitutions, namely, between Position 1 and Position 4 occur extremely rarely and form two polar entities.

DISCUSSION AND CONCLUSIONS

One of the primary issues of concern in this study was whether Russian learners would show any differences in perceiving English monosyllabic words containing 4 possible combinations of tense/lax vowels and fortis/lenis stops in final position.

As predicted, our subjects' perceptual judgements were influenced by specific characteristics of the phonetic realization of the presented stimuli. The easiest to identify were those words whose phonological and phonetic properties do not "contradict" each other, i.e. phonological length of the vowel is "increased" through phonetic lengthening due to the following lenis conconant (Position 1), or phonological shortness of a vowel is made more fully expressed by the shortening effect of the following fortis consonant. The other two word types proved to be more difficult to perceive because of the "contradictory" relationships between the phonological length of tense vowels and their shortening induced by the fortis consonant (Position 2) and the phonological shortness of lax vowels and their lengthening before the lenis consonant.

The intricate interplay of qualitative and quantitative parameters in Positions 2 and 3 results in Russian listeners producing more errors than in the phonetically "more marked" Positions 1 and 4.

On the whole, it may be said that tense vowels demonstrated better identification than lax vowels (84% and 77% respectively), this difference being stronger in the vowels /ɑ: - ʌ/ and /ɔ: - ɔ/ compared with the /i: - i/ vowels.

The data obtained do not show any significant difference in the perception of fortis versus lenis consonants.

Analysis of perceptual errors has shown that Russsian listeners tend to confuse words that differ in one feature only. The vast majority of substitutions involves one feature whereas two-feature errors occur much less frequently.

Among the one-feature confusions one type of error is most widely spread, namely, the substitution of the fortis for the final lenis preceded by a tense vowel. This confusion could be predicted since Russian does not allow for voiced stops in word-final position.

It is interesting to note that there are only a few instances of true indiscrimination between pairs of words where each one is substituted by the other in the approximately the same number of cases, for example, "cod" and "cot" or "sit" and "seat".

Two-feature substitutions occur rather seldom. Of particular interest is the fact that two-feature confusions of a certain type *do* occur. These are the substitutions in the domain of Positions 2 and 3, involving both directions. The other words differing in 2 features (Positions1 and 4) hardly ever get confused.

Summing up, we would like to say that the present study has given some evidence concerning perceptual abilities of Russian learners of English. The phonetic component may be said to be more involved in perception than the phonemic level, the latter being obscured by the phonological models of their mother tongue.

REFERENCES

[1] Chen, M. (1970), Vowel duration variation as a function of the consonantal environment. Phonetica, vol. 22, pp. 129-159.

[2] Barry, W.T. (1983), Perception and production of English Vowels by German Learners: Instrumental-phonetic support in language Teaching. Phonetica, vol. 46, pp. 155-168.

[3] House, A.S., Fairbanks, G. (1953), The Influence of consonant environment upon the secondary acoustical characteristics of vowels. JASA, vol. 25, pp. 105-113.

[4] Kukolshchikova, L.E. (1981), English Vowel Length Revisited. In: Phonetics and Psychology of Speech, 3 (in Russian). Ivanovo, pp. 92-101.

[5] House, A.S. (1961), On vowel duration in English. JASA, vol. 33, pp. 1174-1178.

GESTURAL ECONOMY

Ian Maddieson
University of California, Los Angeles, USA

ABSTRACT

This paper outlines a theory of gestural economy in language structure, with illustration partly drawn from studies of Ewe sounds using electromagnetic articulography and video. It argues that languages tend to be economical both in the number and nature of the gestures used to construct their inventory of contrastive sounds. Tests of this theory are provided by complex consonants and claims of 'polarization' of contrast.

INTRODUCTION

As is well-known, languages show a tendency to construct their inventory of contrastive sounds in a way that is at least partially symmetrical. For example, a language with the stops /p, t, k/ is far more likely to also have /b, d, g/ than to have /ɖ, ɟ, ɢ/. If sounds are regarded as composed of features, this tendency can be expressed as maximum exploitation of compatible feature combinations. The number of features needed to form a given number of contrasts is thus economized.

This paper argues that a similar pattern can be seen in the articulatory organization of the sounds of a language. That is, there is an analogous tendency to be economical in the number and nature of the distinct articulatory gestures used to construct an inventory of contrastive sounds, and it is this (rather than a more abstract featural analysis) that underlies the observed system symmetry. Moreover, this tendency can be seen as an aspect of a more general principle that can be given the name 'Gestural Economy'.

There are three principal strands to the argument in support of this overall view. First, there is the well-known evidence that languages as a whole favor certain articulatory positions and movements, which are by-and-large those that are more efficient (i.e. acoustically effective) and involve less extreme movements. Second, within a language a given articulatory gesture is often exploited for several distinct segments, for example, nasals and stops usually occur at the same places of articulation and complex segments are built up out of gestures used in simple ones. Third, articulations are not generally displaced from the 'economical' positions or otherwise modified when a language includes further contrasts at nearby places. That is, evidence for systematic use of polarization strategies is lacking.

Only a very brief review of the first point will be provided. The second point is supported by a demonstration that the labial and velar gestures in simple bilabial and velar stops are largely similar to those in labial-velar stops in Ewe. The third point will be supported by showing that one of the best-known hypothesized polarization effects is spurious: labio-dental fricatives in Ewe do not ordinarily involve use of an 'enhancing' elevation of the upper lip in these segments.

What is meant by a gesture?

Before proceeding to any further discussion, it may be useful to characterize what is meant by a gesture in the present context. This term is not intended to refer to a primitive element in the organization of phonology (as in Articulatory Phonology [1]), nor to an articulatory invariant. Here, it simply refers to a typical movement trajectory for a given articulatory subsystem in realizing a given phonetic contrast, bearing in mind the initial conditions for the start of the gesture, anticipation of the following context, and any competing demands of other simultaneously specified aspects of the phonetic element of which the gesture is a component. It is thus a recasting of the traditional phonetic notion 'place of articulation' in dynamic terms and with the focus on the properties of the movements of active articulators as much as on the sites at which constrictions are formed.

INVENTORY STRUCTURE

Cross-linguistic studies of segment inventories show that languages tend to include many of the same segments [2]. The stops /p, t, k, b, d, g/, the fricatives /f, s, ʃ/, and the nasals /m, n, ɲ, ŋ/ are more common than other segments of their respective classes. Moreover, languages show a strong tendency to have small 'families' of sounds that share common articulatory positions. This already emerges from the listing of common sounds above, where the sets /p, b, m/, /t, d, n/ and /k, g, ŋ/ share - in traditional phonetic terms - the same place of articulation.

These commonalities are part of the motivation for the proposal of gestural economy. There is good reason to believe that the commonly found articulatory gestures are more frequent since they are in themselves efficient and economical, but further 'economy' is achieved by re-using the same gesture in a variety of segments (even if the gesture is an inherently less economical one), and by resisting uneconomical modifications that might be made in the interests of generating larger acoustic distinctions between competing sounds. The remainder of the paper will illustrate these two points using simple and complex stops as an example of re-use of common gestural patterns, and labio-dental fricatives as an example of the absence of modification.

SIMPLE AND COMPLEX STOPS

Ewe, a language spoken in Ghana and Togo, is among those with the labial-velar stops /k͡p, g͡b/. The component gestures of these labial-velar stops are very similar to those in simple bilabial and velar stops, as discussed in some detail in [3]. In that paper, evidence for the similarity of the gestures in doubly- and singly-articulated stops was illustrated with data from one speaker in an experiment using electromagnetic articulography [4]. Data from a second speaker is presented in Figures 1-3.

Figure 1 shows the time course of the vertical movement of the lower lip in the word /apaa/ *'job'*. In this figure and the next two, the movement data have been converted to standard scores so that they can be plotted on the same scale and with the same origin. Release of the consonant closure, determined from the acoustic record of the utterances, is at 300 ms. This point is used as the line-up point for aligning repetitions. Each of the figures represents the mean of ten repetitions.

Figure 2 shows the vertical movement of a point on the back of the tongue during the plain velar stop in the word /aka/ *'charcoal'*.

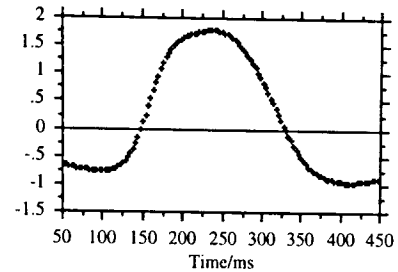


Figure 1. Normalized mean vertical movement of the lower lip in /apaa/

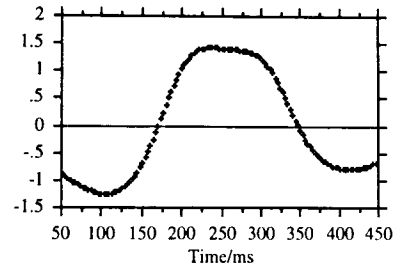


Figure 2. Normalized mean vertical movement of the tongue back in /aka/.

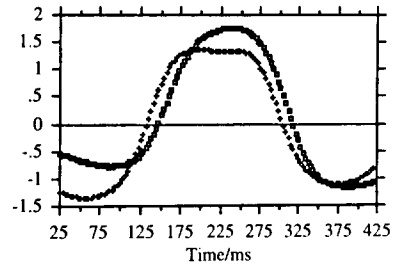


Figure 3. Normalized mean vertical movement of the lower lip and tongue back in /akpa/.

The corresponding movements of both the lower lip and the tongue back during the word /akpa/ *'too much'* are shown in Figure 3. The velar gesture, plotted with small crosses, leads the labial one (small squares) by a few milliseconds, but both gestures are in all salient particulars like those in the simple stops /p/ and /k/. The movements in the doubly-articulated stop have very comparable time courses, very similar shapes, and very similar amplitudes to the movement of the same articulator in a simple stop (amplitude

cannot be read off the normalized plots shown here, but is very comparable in the unnormalized data.) This, of course, need not be the case, and some differences, especially in time course, might have been expected.

Some differences between the oral gestures in voiced and voiceless simple stops at the same place, and between the simple stops and the components in doubly-articulated stops were indeed observed and reported in [3], but with one exception these can be accounted for as contextual effects, due to the demands of other specified aspects of these segments. The exception concerns a backward movement of the tongue body in doubly-articulated stops that is absent in simple velars. Its explanation remains undetermined, but it could be an aerodynamically-induced consequence of a double closure in the oral tract.

Apart from this detail, Ewe doubly-articulated labial-velar stops appear to be made in the simplest way possible – by combining the well-rehearsed movements that are used in simple labial and velar stops. It is not just that labial-velar stops employ two places of articulation that are used elsewhere in the language; they are constructed of the same specific gestures used elsewhere. We hypothesize that languages take maximal advantage of such opportunities for limiting the number of distinct gestures employed, as part of a general preference for gestural economy.

ABSENCE OF POLARIZATION

Ewe is also known as one of the relatively small number of languages with a contrast between [ɸ, β] and [f, v]. It has been claimed [5] that Ewe speakers (and speakers of other languages in the same area with [ɸ, β] and [f, v]) 'enhance' the bilabial/labio-dental contrast among fricatives by using an active raising gesture of the upper lip in the production of the labio-dentals. According to this view, the structure of the set of phonologically significant distinctions in the language has a direct influence on the production of a sound type – a labio-dental fricative – that is among those that are the most highly favored in the world's languages [2].

It seems likely that labio-dental fricatives are favored because this is an optimal place for creating fricatives. It requires precise positioning of only one active articulator rather than two as for a bilabial, and a relatively small movement compared to, say, a linguo-labial or interdental. Labio-dentals are also acoustically readily distinct from all fricatives produced further back – except perhaps [θ].

From a gestural economy perspective, these virtues would be expected to be retained, rather than disturbed because of a contrast with less economical sounds. The articulatory target might in such a case be more precisely defined, constraining the variability in order to protect the contrast, but that is all.

The two Ewe speakers' productions of bilabial and labio-dental fricatives were also investigated using electromagnetic articulography. These speakers showed no upward movement of the upper lip for [f, v]. The upper lip in words such as /eve/ *'two'* remained in the same position as in words like /eke/ *'sand'* [6]. The upper lip lowers quite substantially for [ɸ, β], resulting in a visibly higher lip position for the labio-dentals than for the bilabials. However, this is not due to raising the upper lip in the labio-dentals.

In order to study this question in greater depth, 17 additional Ewe speakers were videotaped saying words contrasting bilabial and labio-dental fricatives, and words containing velar stops in the same vowel environments. In addition, a videotape made earlier of another speaker was analyzed. Both frontal and lateral views of the lips were examined on a frame by frame basis.

A population of 20 Ewe speakers (all from the northern part of the Aŋlo dialect area, where the vowel /e/ is pronounced as a mid front vowel [ɐ] rather than as [ə]) was thus examined. Of these, two show some clear raising of the upper lip in labio-dentals, and two others show some smaller adjustment of the upper lip position either forward or upward. More typical articulations are illustrated in Figures 4 and 5. These figures are digitized frames from the videotape of one of the speakers and show the culminating phase of the word-medial consonants viewed from the side. Both figures show the lip position in /afa/ *'half'* on the left of the figure. For this sound the upper teeth are not completely covered by the upper lip but the lip is not lifted out of the way

to any degree. Figure 4 compares this lip position with that in /aɸa/ *'shout'*. For [ɸ] the upper lip is lowered and drawn inwards to meet the lower lip; it entirely covers the upper teeth. Figure 5 shows that in [f] the upper lip is in a position almost identical to that in the velar stop of /aka/ *'charcoal'* on the right of the figure. The angle of the lip profile below the nose is the same for these two sounds. (Note that a small distortion is introduced just below the superimposed time-coding on the video-tape. This must be ignored in making the comparison.)

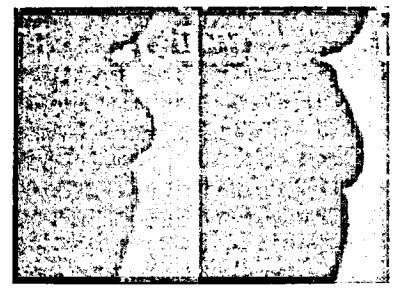


Figure 4. Position of the lips at the center of the consonants in /afa/ 'half' (left) and in /aɸa/ 'shout' (right).

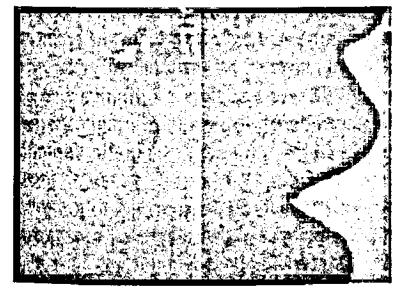


Figure 5. Position of the lips at the center of the consonants in /afa/ 'half' (left) and /aka/ 'charcoal' (right).

Most of the Ewe speakers do not raise the upper lip to produce labio-dentals, but a few do. To determine if this is greater than the cross-speaker variability that one might find in another language without a bilabial/labio-dental contrast, 20 speakers of Sɛle (Santrokofi) were also examined on video-tape. This language, spoken by a people who are neighbors of the Ewe, has only one labial fricative of any kind, [f]. Of this group, two showed a clear raising of the upper lip during [f], three others showed some raising or fronting. Because a more extensive wordlist was taped with the Sɛle speakers, it was also possible to note that the speakers who tended to raise the upper lip for [f] often had a rather similar gesture with certain other consonants, such as [s] and [ɲ].

These data suggest that the occurrence of a raising gesture for labio-dental fricatives is not in any way associated with the presence in the same language of a contrasting bilabial place of articulation for fricatives. Labio-dentals are typically produced in the same way – without an added upper lip gesture – regardless of inventory structure.

SUMMARY

This paper has suggested that some patterns of linguistic structure can be attributed to a principle of gestural economy. Support for this view can be demonstrated both by language-internal comparison across different segments, and cross-linguistically by comparing production of similar segments in differently structured inventories.

ACKNOWLEDGEMENTS

Work supported in part by the National Science Foundation in the US and the US Information Agency in Ghana.

REFERENCES

[1] Browman, C. P. & L. Goldstein (1992), "Articulatory Phonology: an overview", *Phonetica*, 49: 155-80.

[2] Maddieson, I (1984), *Patterns of Sounds*, Cambridge: C. U. P.

[3] Maddieson, I. (1993), "Investigating Ewe articulations with electromagnetic articulography", *Forschungsberichte des Instituts für Phonetik und Kommunikation der Universität München* 31: 181-214. (Also see *UCLA Working Papers in Phonetics* 85, 22-53.)

[4] Perkell, J. S., M. Cohen, M. Svirsky, M. Matthies, I. Garabieta & M. Jackson (1992), "Electromagnetic Midsagittal Articulometer (EMMA) systems for transducing speech articulatory movements", *JASA*, 92: 3078-96.

[5] Ladefoged, P. (1993), *A Course in Phonetics* (3rd Edition), New York: Harcourt Brace Jovanovich.

[6] Ladefoged, P. & I. Maddieson, (1995), *Sounds of the World's Languages*, Oxford: Blackwells.

VARIABILITY IN GLOTTALIZATION OF WORD ONSET VOWELS IN AMERICAN ENGLISH

Laura C. Dilley and Stefanie Shattuck-Hufnagel
Research Laboratory of Electronics, MIT, Cambridge, MA 02139, USA

ABSTRACT

American English glottalization of word-onset vowels at the beginning of a new intonational phrase or at a pitch accent holds across speakers and speaking styles, although individuals differ in glottalization rates and in sensitivity to prosodic elements, and manifest glottalization in a number of ways.

1. INTRODUCTION

Speakers of American English often glottalize in certain locations, e.g. at the ends of phrases ("final creak"), at a final voiceless stop (glottalized /t/) and at a word-onset vowel, as in "able" or "alone". For word-onset vowels, the likelihood of glottalization is increased significantly in certain prosodic contexts, e.g. when the vowel-initial word begins a new intonational phrase or is pitch accented [7]. For reduced vowels, glottalization is more likely when the new phrase is a full rather than an intermediate intonational phrase; for full vowels, it is more likely when the pitch accent occurs on the target syllable rather than later in the word [2].

Although prosodic boundaries and prominences constrain the glottalization of word-initial vowels in similar ways across speakers, there is also substantial variation among speakers [3]. For example, for 4 radio news broadcasters, overall glottalization rates were 41%, 37%, 22% and 7%.

This striking rate variation raises several questions. Do the differences arise in part from the use of different texts? Do the findings hold across speaking styles? Can individual differences in onset-vowel glottalization be related to other aspects of glottalization behavior? This paper compares glottalization rates for word-initial vowels for FM radio news speakers producing the same texts, for non-professional speakers reading isolated sentences, and for speakers producing spontaneous speech.

2. FM RADIO NEWS SPEAKERS

Since the use of different texts may partially account for the range of glottalization rates in Dilley et al. [2], we analysed a corpus in which all speakers produce the same texts, for evidence that a) speakers differ in their prosodic interpretation of those texts, and b) these differences might contribute to contrasts in overall glottalization rate. Four news stories were originally broadcast by one speaker and later read in the lab by all the speakers in FM news style. Renditions of all 4 stories were available for 5 speakers (3 female, 2 male) including the original 4.

2.1 Database and Analysis. Details of the BU FM Radio News corpus are described in Ostendorf et al. [6]. Briefly, the speech was recorded in the studio during broadcast, orthographically transcribed, phonetically aligned and labelled for part of speech and for prosody, using the ToBI transcription system [10], [8]. The ToBI (TOnes and Break Indices) system marks 5 levels of prosodic constituent boundaries, and, for each intonational phrase, location (by syllable) and type of pitch accents, location and type of phrase accent, location and type of boundary tone, and highest prominence-related F0. For this study, additional labels were added by hand to each vowel that began a word: phonetic lexical stress (Full Vowel vs. Reduced Vowel), and glottalization (+Glot vs. -Glot). The criteria for glottalization were two-fold: a perceptual impression of glottalization, and a marker in the wave form, usually an irregularity in pitch period duration (or a dip in F0), but occasionally in wave

form shape (e.g. diplophonia). Tokens which did not satisfy both of these criteria were labelled as -Glot, or as Questionable Glot e.g. if the boundary between the target vowel and a preceding glottalized segment was unclear, or only one of the two criteria was met. The four stories contained 573 vowel-initial words in about 5 minutes of speech per speaker. We noted the overall glottalization rate for word-initial vowels for each speaker, and analysed the effect of several types of prosodic environment. Results for tokens preceded by a syllable that contained a glottalized segment (often the result of phrase-final creak in the preceding phrase), and/or by a pause were analysed separately. Removed from the analysis were tokens with questionable phonetic stress (2% of total) and tokens with questionable glottalization (2%).

2.2 Results and Discussion. Rates of glottalization for word-initial vowels preceded by a pause > 50 ms or earlier glottalization were 94-100%, indicating that these factors are strongly associated with high glottalization rates. The number of tokens with such preceding contexts varied from a high of 88 (speaker f2) to a low of 19 (speaker m1), illustrating the fact that different speakers produced different distributions of pause and/or word-final glottalization for the same text.

Table 1 shows results for all 5 speakers for vowel tokens not preceded by a pause or creaky segment. Rates are shown separately for all tokens, for pitch accent and phrase-intial contexts, and for the remaining contexts.

Table 1. Glottalization rates overall (a), for pitch accent and phrase-onset contexts (b), and for the remaining contexts (c) for 5 FM radio news speakers.

%	f1	f2	f3	m1	m2
(a)	32	34	29	13	16
(b)	67	58	62	30	33
(c)	8	15	6	2	2

Individual overall glottalization rates for word-onset vowels range from 34% to 13%. In addition, glottalization is more likely at prosodic phrase boundaries and pitch accents, even for speakers m1 and m2, who have low rates of glottalization. ($p \leq .0001$)

Additional examples of speaker differences hint at a gender effect. For example, the 3 female speakers glottalized non-phrase-initial Full Vowels in pitch accented words significantly more often than in non-pitch-accented words ($p \leq .01$), but the 2 male speakers did not. Moreover, the 3 female speakers in Table 1 have higher rates (34%, 32%, 29%) than the 2 males (16% and 13%). To determine whether similar effects arise for other speakers, we turned to a speech database of isolated sentences read by non-professional speakers in the lab.

3. NON–PROFESSIONAL SPEECH

3.1 Database and Analysis. The LEX database consists of isolated sentences read aloud in the lab by 4 non-professional speakers (2 female, 2 male) and digitized. For the present study, we selected the sentences that contained word-onset vowels, and labelled the prosody and glottalization as above. The overall rates and the effect of prosodically significant contexts are shown in Table 2.

Table 2. Glottalization rates overall (a), for pitch accent and phrase onset contexts (b), and for the remaining contexts (c), for 4 non-professional speakers.

%	f-j	f-s	m-k	m-m
(a)	29	5	<1	13
(b)	67	12	3	43
(c)	6	**0**	**0**	**0**

3.3 Results and Discussion. Like the FM newscasters, these non-professional speakers show a range of overall glottalization rates, from less than 1% to 29%, and a tendency for Phrase Initial and Pitch Accent contexts to elicit more glottalization. Speaker m-k was the exception: his overall glottalization rate was so low that distinctions between prosodic contexts did not emerge. The pattern of lower glottalization rates for male speakers observed in the radio news speech is called into question: female speaker f-s showed one of the

lowest rates of glottalization, both overall (5%) and for prosodically significant contexts (12%), and both her rates were lower than those for male speaker m-m (13% and 43%), indicating that individual variation makes it difficult to assess gender-related differences with a small number of speakers.

4. SPONTANEOUS SPEECH

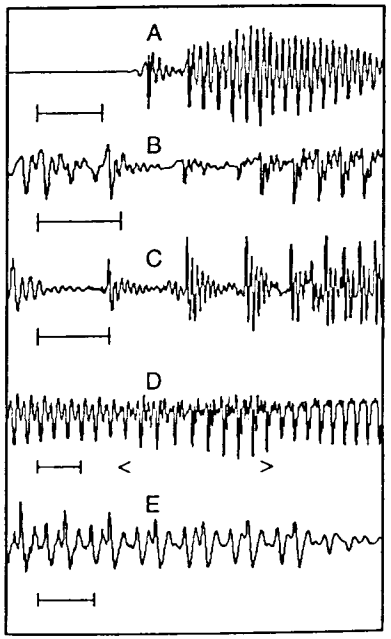
4.1 Database and Analysis. To determine whether the influence of prosodic context on glottalization rates is also found in spontaneous speech, we analyzed a set of utterances from the ATIS corpus, collected from travel agents enacting the task of making airline reservations using spoken language to interact with a computer [12]. We examined a subset of 155 utterances spoken by many different speakers, each containing a disfluency. The digitized utterances were aligned and labelled as above. Since each speaker produced just a few utterances, only descriptive results are given here. Tokens with a previous pause or glottalized syllable were removed, leaving 138 word-onset vowels for analysis.

4.2 Results and Discussion. The overall rate of glottalization for this corpus was 43%, with a greater likelihood of glottalization for pitch accented and/or phrase initial vowels (64% of 78 tokens) vs. other prosodic contexts (17% of 60 tokens). An interesting aspect of these utterances was the finding that, for reduced vowels, tokens with a pitch accent later in the word were more likely to be glottalized than tokens with no pitch accent on the word (43% vs. 5%). The effect of a pitch accent later in the word was also observed for a single FM radio news speaker by Dilley et al. [3], and suggests that the effects of an accent may extend beyond the boundaries of its immediate syllable (see alsoTurk [11]).

5. WAVE FORM SHAPES

A critical aspect of any study of glottalization rates is the decision about what to categorize as glottalization. Our two-part criterion, perceptual salience and pitch period irregularity, labelled a variety of waveform shapes as +Glot; some examples are shown in Fig. 1.

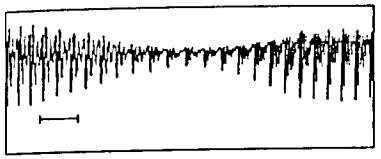
Figure 1. Examples of word-onset vowels heard and labelled as glottalized (FM radio speakers). A: glottal stop, B: general irregularity, C: ecponentially-decaying pulses, D: F0 dip and amplitude change (< > indicates region of lower F0), E: diplophonia. Indicator bar = 20 msec.



Excluded from this category was a kind of vowel onset which did not sound glottalized, yet gave the impression of a boundary or onset marker. This was most common for FM radio news speaker m1, where it was often associated with a dip in amplitude, without other noticeable irregularity in the wave form, e.g. Fig. 2. Houde and Hillenbrand [4] have reported that a dip in amplitude is a sufficient cue to elicit the perception of glottal onsets for the vowels in a synthesized version of the exclamation "oh-oh".

Another speaker, m-k, a non-professional, marked onsets in a perceptually salient way such that they sounded "breathy". We did not observe

Figure 2. Example of a dip in amplitude, heard as onset-like but not as glottalized (speaker m2)



irregularity in the period or shape of the waveform in these cases, but noise was present in the signal; moreover, these tokens had an /h/-like distribution of energy across frequencies. Tokens which were "breathy" were labelled -Glot. The fact that m-k appears to employ this marking strategy quite often instead of glottalizing contributes to his low overall rate of glottalization.

These observations suggest that a wide variety of signal characteristics and perceptually salient phenomena can occur at word-onset vowels, underlining the importance of a better understanding of the glottal mechanisms involved.

5. CONCLUSIONS

Analysis of word-onset vowel glottalization rates in American English, and their variation across individual speakers, prosodic contexts and speaking styles, shows that a) individuals vary substantially in overall glottalization rate as well as in use of prosodic structure, and b) a variety of signal shapes are perceived as glottalized in these vowels. Nevertheless, these preliminary analyses support the view that prosodic structure plays an important role in determining where vowel glottalization will occur. Somewhat similar findings have been reported for German by Kohler (1994). Further work will determine whether glottalization can provide useful cues to prosodic structure, not only as phrase-final creak but also as a marker for phrase onset and pitch accent, and whether appropriate use of glottalization can increase the naturalness and comprehensibility of synthesized speech.

ACKNOWLEDGEMENTS

This research was supported by NSF, NIH and ARPA, and by the MIT Undergraduate Research Opportunity Program. The original analysis of glottalization in FM radio news speech was carried out with Mari Ostendorf; the BU Radio News Corpus was constructed in collaboration with Mari Ostendorf and Patti Price; and the LEX database with Ken Stevens and Sharon Manuel. We gratefully acknowledge help provided by Melanie Matthies.

REFERENCES

[1] Beckman, M. and Pierrehumbert, J. (1986) Intonational structure in Japanese and English, *Phonology Yearbook* 3, 15-70

[2] Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M. (1994) Prosodic constraints on glottalization of vowel-initial syllables in American English, *JASA* 94 (Pt 2), 2978

[3] Dilley, L., Shattuck-Hufnagel, S. and Ostendorf, M. (submitted) Glottalization of vowel-initial syllables as a function of prosodic structure

[4] Houde, R.A. and Hillenbrand, J. (1994) The role of voice pitch in the perception of glottal stops, *JASA* 95 (S1), 2872

[5] Kohler, K. J. (1994) Glottal stops and glottalization in German, *Phonetica* 51, 38-51

[6] Ostendorf, M., Price, P. and Shattuck-Hufnagel, S. (1995) *The Boston University Radio News Corpus*, Boston University ECS Enginerring Report No. ECS-95-001

[7] Pierrehumbert, J. and Talkin, D. (1992) Lenition of /h/ and glottal stop, in *Papers in Laboratory Phonology II* (G. Doherty and D.R. Ladd, eds.), 90-117. Cambridge University Press.

[8] Pitrelli, J. F., Beckman, M.E. and Hirschberg, J. (1994) Evaluation of prosodic transcription labelling reliability in the ToBI framework, *Proc. Int. Conf. on Spoken Language Processing*,

[9] Shattuck-Hufnagel, S., Ostendorf, M. and Ross, K. (1995) Stress shift and early pitch accent placement in lexical items in American English, *J. Phonetics* 22, 357-388

[10] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. (1992), TOBI: A standard for labelling English prosody, *Proc. Int. Conf. Spoken Lang. Processing* 867-870

[11] Turk, A. and Sawusch, J.R. (forthcoming) The domain of the durational effects of accent, MIT Speech Group Working Papers 10

[12] Multi-Site Data Collection for a Spoken Language Corpus, MADCOW, Proc. DARPA Speech and Nat. Lang. Workshop Feb 1992 p. 7

THE ALARYNGEAL VOICE SOURCE AS ANALYSED BY VIDEOFLUOROSCOPY, FIBERENDOSCOPY, AND PERCEPTUAL - ACOUSTIC ASSESSMENT

Britta Hammarberg [1,2] and Lennart Nord [2]

[1] Dept of Logopedics and Phoniatrics, Karolinska Inst, Huddinge Univ Hospital, Huddinge, Sweden, and [2] Dept of Speech Communication and Music Acoustics, KTH, Stockholm, Sweden.

ABSTRACT

Vibratory characteristics of the pharyngo-esophageal (PE) segment, which constitutes the laryngectomee voice source, were related to perceived and acoustic voice qualities, with special emphasis on the voiced-voiceless distinction in stop sounds. Four proficient laryngectomee speakers were included, representing both tracheo-esophageal fistula characterised by either of two strategies: (1) an opening gesture of the back wall of the PE-segment in the fiberoptic speech and esophageal speech. Videofluoroscopic and fiberstroboscopic images of the PE-segment were recorded during phonatory tasks. Results indicated that 84% of the voiced - voiceless word pairs were audibly distinguished. The voiced-voiceless contrast was registration prior to the production of a voiceless stop, and a more forceful upward movement of mucus and barium contrast in the voiceless stops, that was not observed as clearly in the voiced cognate sounds, and (2) a slight prephonatory delay in the closing of the back wall towards the front wall in the initial phase of voiceless stops was observed in the videofluoroscopic registrations. The voiced - voiceless contrasts were automatically confirmed by spectrograms.

INTRODUCTION

In the most commonly used laryngectomee speech techniques the voice source is situated in the upper part of the esophageus, the so-called pharyngo-esophageal (PE) segment. This segment is a multi-layered structure of mucosa and muscle, quite similar to the vocal fold structure. The PE-segment is brought into vibration either on air that has been injected into the esophagus from the mouth, so called esophageal (E) speech, or on pulmonary air that is led into the esophageus via a valve placed in a tracheo-esophageal fistula, so called tracheo-esophageal (TE) speech. Although for both methods the PE-segment is used as voice source, the air volume during phonation is much larger in TE-speech, since the lungs are used. Thus, TE-speakers produce louder and more fluent speech [1,2,3]. As regards consonant intelligibility, however, earlier results indicated that the two speaker groups did not differ significantly from each other, with a mean consonant intelligibility score of 83% for the E-speakers and 86% for the TE-speakers [4].

PURPOSE

The purpose of this study was to relate the vibratory characteristics of the PE-segment to perceived and acoustic voice qualities, with special emphasis on the voiced-voiceless distinction in stop sounds. The results aim at enhancing our knowledge of the physiological and structural characteristics of the alaryngeal voice source and its functional constraints.

The present study is part of an ongoing project that also comprises aspects such as communicative efficiency in background noise, ratings by experienced and naive listeners, and aerodynamic/acoustic measurements.

Table 1. The speakers

	Sex	Speaking technique	Age	Years since op.
No 1	man	tracheo-esophageal	67	2
No 2	man	tracheo-esophageal/esophageal	54	3
No 3	man	esophageal	52	12
No 4	woman	tracheo-esophageal	55	2

METHODS

Speakers

Four speakers were recorded, representing both TE-speech and E-speech (see Table 1 for details). One of the male TE-speakers (No. 2) also mastered E-speech. All subjects were regarded as proficient speakers and had maintained their occupations as military officer, foreman, technician and nurse.

Analyses

Videofluoroscopic images of the PE-segment were recorded during phonatory tasks (and swallowing) in frontal and lateral projections with a rate of 25 pictures per second. Prior to the registration the subject swallowed Barium contrast (Mixobar High Density) to cover the walls in the pharynx and the esophagus. *Audio recordings* of the phonatory tasks were simultaneously made on the same video recorder as was used for the videofluoroscopic registrations and on a separate high quality DAT recorder.

Two of the speakers (Nos. 2 and 3) were also recorded by *videofiberstroboscopy* performing the same phonatory tasks. A fiberoptic laryngoscope (3.5 mm Olympus ENF-P) was connected to a stroboscope (Bruel & Kjaer 4914) and to video equipment. The fiberscope was inserted through one of the nostrils and placed with the tip in the pharynx just above the esophageal entrance.

Acoustic analysis included mean and range of fundamental frequency, spectral characteristics, such as the level of fundamental relative to the level of formants, sound pressure level, and segmental observations. Computer-based analysis programs, developed at the KTH department were used [5,6,7].

Perceptual evaluation by the two authors included listening to voice quality and pitch, and rating consonant intelligibility from the recordings.

Speech Material

Speakers were asked to read aloud a standard text and word pairs containing either a voiced or an unvoiced stop consonant initially or finally: /ka:l, ga:l/, /bank, pank/, /dom, tom/, /bus:, pus:/, /dil:, til:/, /lab:, lap:/, /bu:d, bu:t/, /lok:, log:/, /vi:t, vi:d/, /jø:k, jø:g/.

RESULTS

Voice Source Characteristics

Videofluoroscopic observations showed that in the three male speakers the voice source was situated in the PE-segment, i.e. the bulging semicircular segment in the back wall of the lowest part of the pharynx and the upper part of the esophagus. In the female TE-speaker, however, the voice source seemed to be made up of two structures, a vibrating constriction situated about 2 centimetres below the PE-segment down in the esophagus (a "subsegment") and the PE-segment. The back wall of the PE-segment never seemed to close towards the front wall during phonation, while there was a full closure in the "subsegment." This finding led us to conclude that the lower constriction was the primary voice source.

Fiberendoscopic observations of two of the speakers (Nos. 2 and 3) showed vibrations in the esophageal orifice with vibratory movements from the back wall towards the frontal wall of the PE-segment. The amplitudes of the vibrations were larger in the back wall than in the frontal wall. A clear mucosal wave was also observed, reminiscent of the glottal mucosal wave. Short sequences of regular vibrations of the PE-segment in one of the patients allowed stroboscopic images to be registered. These indicated a pattern of successive

closures and openings of the segment. This subject (No. 2), who mastered both TE-speech and E-speech, was also able to open the PE-segment on command, a manœuvre that he used for prephonatory air-intake in his E-speech.

Perceptual evaluation of the voices showed that the two male TE-speakers (Nos. 1 and 2) had rather rough, strong and low-pitched voices, i.e. "classical" laryngectomee voices, and that the E-speaker (No. 3) had a rather high-pitched and somewhat weak voice, whereas the female TE-speaker (No.4) had a hyperfunctional, strained and weak voice.

Data from the *acoustic* analysis confirm these observations, see Table 2.

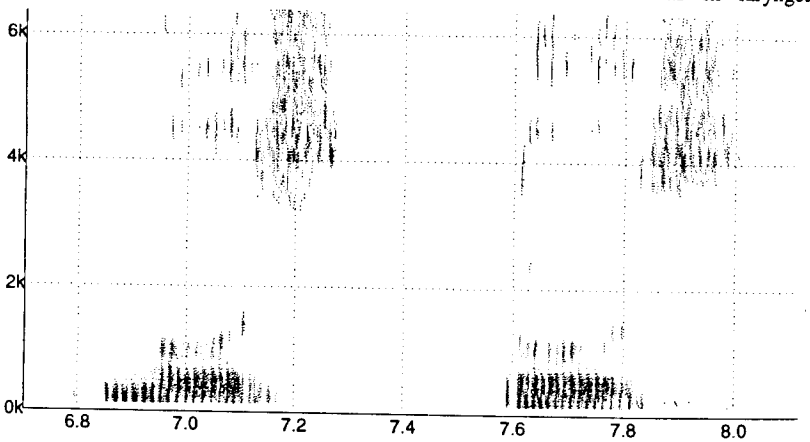


Figure 1. Spectrograms of the voiced-voiceless distinction in the word pair [bes: pes:].

Voiced - Voiceless Distinction

As regards the voiced - voiceless contrast, both *videofluoroscopic* and *fiberoptic* registrations seemed to confirm that there was some kind of gesture difference in the PE-segment when the contrast was heard. There seemed to be two strategies for maintaining the distinction. (1) There was an indication of an opening gesture of the back wall in the fiberoptic registration prior to the production of a voiceless stop which was not seen as clear in the voiced cognate sound. The opening gesture consisted of a lifting of the back wall of the PE-segment. Also, there was a more forceful upward movement of mucus and Barium contrast in the voiceless stops. (2) A second strategy was observed in the videofluoroscopic registration: a slight prephonatory delay in the closing of the back wall towards the front wall in the initial phase of voiceless stops.

As for the *perceptual evaluation* there were audible contrasts between voiced - voiceless stops in about 84% (75-92%) of the word pairs during the videofluoroscopic and fiberscopic recordings. *Acoustic analysis* of voiced - voiceless contrasts in the present study showed that when the distinctions were mastered, they were realized by the same acoustic cues as in laryngeal speech, i.e. voice bar for the voiced stops and occlusion for the voiceless stops, see Fig. 1. Voice onset times were close to what Hirose et al. [8] have found for Japanese alaryngeal speech.

DISCUSSION

In our earlier studies of consonant intelligibility tests, the distinction voiced - voiceless in stops has proved to be difficult for laryngectomees to master [4]. There is evidence, however that the distinction can be mastered by some laryngectomees. In the present study of four proficient tracheo-esophageal/esophageal speakers, about 84% of voiced - voiceless word pairs were audibly distinguished. The contrasts were acoustically confirmed by spectrograms.

What constitutes the voiced - voiceless distinction in alaryngeal speech? From the preliminary findings of the fiberoptic and videofluoroscopic registrations, we observed an opening gesture in the voiceless stops. In some cases this gesture was followed by a more forceful occlusive phase in the voiceless sounds, realised by a larger amount of upgoing mucus and Barium during the registrations. The latter observation might be interpreted as a *fortis - lenis* contrast, which is one of the phonetic cues of this distinction in Swedish. Also in the spectrographic analyses a minimal *aspiration* phase was seen in unvoiced stops, which corresponds to the prolonged opening gesture observed in some voiceless sounds. This is in agreement with Hirose et al. [8], who observed a transient opening of the PE-segment for the production of voiceless consonants.

The speaking rates were found to be within the normal range, which is in accordance with the general view that these speakers were proficient.

ACKNOWLEDGEMENTS

We are grateful to Nick Edsborg, Dept of Radiology, Huddinge Univ. Hospital, for making the videofluoroscopic registrations, and to Per-Åke Lindestad, Dept of Logopedics and Phoniatrics, Huddinge Univ. Hospital, for the fiberendoscopy.

This investigation was supported in part by grants from the Swedish Council for Social Sciences and from the Karolinska institute.

REFERENCES

[1] Robbins J, Fischer H, Blom E, Singer, M (1984): A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *J of Speech and Hearing Disorders*, vol. 49, pp. 202-210.

[2] Perry, A. (1988): Surgical voice restoration following laryngectomy: the tracheo-oesophageal fistula technique (Singer-Blom). *British J of Disorders of Communication*, vol. 23, pp. 23-30.

[3] Hammarberg, B. & Nord, L. (1988): Communicative aspects of laryngectomee speech. Presentation of a project and some preliminary results, *Phoniatric & Logopedic Progress Report*, Huddinge Univ Hospital, vol. 6, pp. 10-27.

[4] Hammarberg, B., Lundström, E. & Nord, L. (1990): Consonant intelligibility in esophageal and tracehoesophageal speech. A progress report. *Phoniatric & Logopedic Progress Report*, Huddinge Univ Hospital, vol. 7, pp. 49-57.

[5] Liljencrants, J. (1988): Spectrogram Program "SPEG". Custom Computer Program. Dept Speech Comm & Music Acoustics, KTH, Stockholm, Sweden.

[6] Carlson, R. (1988): Waveform Editor "MIX". Custom Computer Program. Dept Speech Comm & Music Acoustics, KTH, Stockholm, Sweden.

[7] Ternström, S. (1992): Soundswell - Signal Workstation Software Manual, Ver.3.0., Sounds-well Music Acoustics HB, Sollentuna.

[8] Hirose H, Sawashima, M, Yoshioka, H (1983):Voicing distinction in esophageal speech - perceptual, fiberoptic and acoustic studies. *Ann Bull Research Inst Log Phon*, vol.17, pp.187-199.

Table 2. Acoustic measures for the four speakers.

	Speaking rate, including pauses (syll/sec)	Speaking rate, excluding pauses (syll/sec)	Pauses, in % of total reading time	Mean F0 (Hz)	SPL, 1m (dB)	L0, 1 m (dB)
No. 1/TE	3.3	5.3	38	75	66	42
No. 2/TE	4.1	6.1	33	78	55	41
No. 3/E	3.0	4.7	37	137	54	44
No. 4/TE	3.0	3.9	24	150	43	28

THERAPEUTIC TRAINING OF AN RHD PATIENT WITH PROSODIC DISTURBANCES OF A LINGUISTIC AND AFFECTIVE NATURE

Deborah Günzburger, Research Institute of Language & Speech, Utrecht University, the Netherlands, Laura Roelants & Mariska v. Schothorst

ABSTRACT

The core of this paper is a case study of a right hemisphere disorder (RHD) patient whose improvement in the production of affective prosody is described in the light of a) general issues of differential lateralization of the affective and linguistic aspects of language and b) a specific therapeutic training that involved visual feedback as part of his treatment.

INTRODUCTION

For more than a century (Broca, 1865, Wernicke, 1977, originally published in 1874) it has been known that brain injuries in the left hemisphere can cause impairment in various linguistic skills and consequently the claim has been made that language is a lateralized function of the left hemisphere; the role of the right hemisphere in language has, for a long time, been considered rudimentary. However, over the last two decades it has become increasingly evident that the right hemisphere also has an active role in language processing, in particular in the domain of affective prosody.

The right hemisphere superiority for emotional prosody has been meticulously scrutinized by Ross, Edmondson, Seibert & Chan, 1992, by testing Taiwanese speaking subjects who had incurred infarctions of the right frontoparietal region as documented by tomographic brain scans. The patients were able to use F0 modulations for the production of lexical tones, but showed impairment in their ability to use tone latitude for affective expression. These data are in line with the postulated differential lateralization of linguistic and emotional prosodic aspects in the human brain.

In non-tone languages, of course, non-lexical affective signalling in speech is mainly realized by pitch variation in the form of intonation, with various other factors also having a possible contributing role. Since the **acoustical** properties of linguistic and emotional prosody are identical (e.g. Lieberman, 1967) the issue of different localization in the brain is particularly interesting but also extremely complex.

Various data of RHD patients indicate the importance of the exact lesion site (Shapiro & Danly, 1985): patients with right anterior brain damage had less pitch variation in their speech and a more restricted intonational range as compared with normal speakers. Patients with right central brain damage displayed a similar pattern with, in addition, a lower mean F0 level. This observation applies to the emotional and the propositional domain. Patients with right posterior damage had a higher F0 level and more pitch variation than either right anterior -, right central -, left posterior brain damage patients or normal control subjects.

In addition, Behrens (1988) has proposed the so called functional lateralization, a model that applies also to non-tone languages. This model assumes that the right hemisphere is superior for the processing of emotional prosody but not for linguistic prosody. In Behrens' (1968) study, RHD patients' ability to convey

purely linguistic stress appeared preserved, whereas use of emotional prosodic cues was severely impaired. It should be noted that the combinational influence of lesion site and functional lateralization adds to the complexity of the issue.

To gain some insight into this subject matter, and stimulated by the clinical background of one of the authors, an experiment was carried out with a clearly therapeutic goal in mind. This experiment, a case study, will now be described in comprehensive terms; for details see also Roelants & van Schothorst, 1993.

METHOD

The subject was a 39 years old male RHD patient. His brain damage was due to a right frontal subdural hematoma. Neurological diagnosis also showed spastic hemiparesis. Patient's speech was monotonous with irregularities in loudness and rhythm. His perception of prosodic cues seemed reduced. Since his visual perception was intact, treatment with a visual feedback method was chosen. We used the so called "Speech Viewer" developed by IBM, which produces an on-line visual display of pitch and loudness variations. To the best of our knowledge there is no description of prosodic training with visual feedback for RHD patients, whereas such programmes have been widely used and are well documented for the deaf and hearing impaired. The visualization of prosodic cues can:

- enhance patient's general motivation for the training programme,
- facilitate the learning process as an additional sensory input modality and
- stimulate a possibly reduced "self monitoring" function.

The influence of visual feedback on the self monitoring task in connection with RHD patients is discussed in greater detail in Roelants and Van Schothorst, 1993 and does not fit within the scope of the present paper.

The above mentioned functional lateralization model assumes a right hemispheric superiority for emotional prosody processing and a left hemispheric superiority for linguistic prosody processing. Therefore the training programme we constructed was focused on a well structured increase of emotional pitch variation. As a first therapeutical attempt it was decided to train the least damaged, left hemispherically controlled, most linguistically significant aspects of prosody.

Expectations were that

1. all trained prosodic aspects would improve with a maximum increase in improvement due to emotional prosodic training, moderate improvement due to sentence intonation training and relatively little improvement due to prosodic training based on contrastive accent (see also our section on speech material).
2. a positive transfer would take place from speech produced during training sessions to utterances produced in more natural conditions.

Speech produced by the patient before, during and after training sessions was perceptually evaluated by a panel of listeners in terms of naturalness and related to some acoustic features.

The speech material of the training programme consisted of the following parts (conditions):

1. contrastive lexical accent: a certain sentence was given and answers to ensuing questions were elicited, e.g. "father was at the office yesterday"; "who was at the office yesterday?"; "where was father yesterday?"
2. sentential intonation: training of interrogative, declarative and imperative sentences.
3. emotional tone: expression of a given emotion based on situational information. E.g: after months of having been looking for work, you are selected for the desired job. Extremely happy, you tell your best friend: "……!"

These prosodic conditions were each trained three times a week for two consecutive weeks. Purely imitational exercises were followed by pseudo-

spontaneous role play between patient and therapist. All training was enhanced by on-line visual feedback using the modules 'pitch' and 'loudness' of the 'Speech Viewer'.

Before and after training transfer of the prosodic aspects was tested by means of

4. reading aloud a text with a high occurrence of direct speech and
5. a dialogue between patient and therapist.

PERCEPTUAL EVALUATION

Before, during and after training, recordings were made of nine sentences per condition for the purpose of perceptual evaluation. In order to evaluate the so called transfer, additional recordings were made of two times nine sentences before and after training (see also Table I).

For the actual listening test sentence pairs were prepared. A pair consisted of two sentences in any combination of recordings (before, during, after) within a condition. Pairs were presented in pseudo-randomized order to a panel of 32 speech-therapy students who were instructed to pay special attention to prosodic cues when judging the sentences as to naturalness. Comparison scores were given on the second sentence of each pair on a 5-point scale.

Raw data were subsequently processed and analysed by means of the Scheffé (1952) method. Ensuing results are expressed in so called preference values (range: 0 - 2) and showed an improvement in all three prosodic conditions and the two transfer conditions. Table I indicates whether improvement is significant. Data are pooled over 32 listeners.

Table I: scheme of perceptual evaluation with indication of significance (= significant on $p < .05$)*

condition	1st.rec.	2nd.rec.	sign.	pref.val.
contr.accent	before	after	*	.4
	before	during	*	.36
	during	after	-	.04
sent.inton.	before	after	*	.43
	before	during	-	.13
	during	after	*	.3
emotion.tone	before	after	*	1.1
	before	during	*	.83
	during	after	*	.28
reading	before	after	*	.54
dialogue	before	after	*	.2

When trying to relate these perceptual findings to some acoustic characteristics the following observations can be made:

1. Improvement in judgments in the contrastive accent condition can be related to the fact that in the course of the training programme the lexical item in question is realized with higher F0, greater amplitude and longer duration.
2. Higher naturalness judgments in the sentence intonation condition are due to a better command of a number of prosodic rules: sentence terminal F0 decrease instead of increase for declarative sentences, increase in amplitude for imperative sentences and overall decrease of speech monotony.
3. Inspection of data on the emotional tone condition indicates a bipartition of emotions into 'restrained' emotions like shy, anxious or disappointed and 'effusive' emotions like happy, angry or surprised. Whereas before training there was little or no distinction between the two types of emotions, this was clearly different after training: effusive emotions were characterised by a relatively high average F0, greater amount of F0 variation and higher average amplitude value; for restrained emotions, measurements indicated a relatively low average F0, less F0 variation and lower amplitude values.

CONCLUSION

As was expected and can be seen in Table I, greatest improvement took place in the emotional tone condition. Overall degree of improvement of sentence intonation and contrastive accent is of a similar magnitude, the latter reaching its maximum value at an earlier stage in the training programme than the former. It is probable, therefore, that follow-up training of sentence intonation and emotional tone would further enhance positive results of the present programme.

We can state that training emotion in prosody with a visual feedback method turned out effective in the case of our RHD patient. Since it concerns an N=1 study, further research with more subjects and a matched group of nonneurological speakers is needed. It should be kept in mind, however, that this kind of study cannot provide an answer to the more fundamental question as formulated by Bates (1994, personal communication): We do not know with any certainty whether localized language deficits due to brain injury can be improved as a result of domain specific (partial) recovery or that it is successful reorganization of regional specialization (stimulated by appropriate therapy) that is responsible for the observed improvement.

REFERENCES

[1] Broca, P. 1865, Du siège de la faculté du langage articulé. *Bulletin de la societé d'Anthropologie*, 6: 377-393.
[2] Wernicke, C. 1977, *Der aphasische Symptomencomplex. Eine psychologische Studie auf anatomischer Basis*. In Wernicke's works on aphasia. Paris, Mouton.
[3] Ross, E.D., Edmondson, J.A., Seibert, G.B., & Chan, J.-L. 1992, Affective exploitation of tone in Taiwanese: an acoustical study of "tone latitude". *Journal of Phonetics*, 20: 441-456.
[4] Lieberman, P. 1967, *Intonation, Perception and Language*, Cambridge, MIT Press.
[5] Shapiro, B.E. & Danly, M. 1985, The Role of the right hemisphere in the control of speech prosody in propositional and affective contexts. *Brain and Language*, 25: 19-36.
[6] Behrens, S.J. 1988, The role of the right hemisphere in the production of linguistic stress. *Brain and Language*, 33: 104-127.
[7] Roelants, L.M. & Schothorst, M.M. van 1993, Prosodische training met visuele terugmelding bij een patiënt met rechter hersenhelft beschadiging. *Stem-, Spraak- en Taalpathologie*, 4: 257-270.
[8] Sheffé, H. 1952, An analysis of variance for paired comparison. *Journal of the Statistical Association of America*, 47: 381-400.

A METHOD FOR THE FIELD EXAMINATION AND FOLLOW-UP OF VOICE THERAPY IN PROFESSIONAL VOICE USERS WITH VOCAL FATIGUE

Leena Rantala[1], Kari Haataja[2], Erkki Vilkman[2]
[1]Department of Finnish, Saami and Logopedics, and [2]Department of Otolaryngology and Phoniatrics, University of Oulu, Oulu, Finland

ABSTRACT

Speech data of ten female teachers was collected from their first and last lessons of a working day by means of a portable DAT-recorder. The mean fundamental frequency (F0) at the beginning of the first lesson was significantly lower than that of the last one (p=0.059). Although the within-teacher variation of the F0 was very idiosyncratic, trends seen in the changes of pitch during the lessons can be ralated to the symptoms of vocal fatigue.

INTRODUCTION

Most of the studies that have been made about vocal fatigue are either questionnaire surveys or laboratory examinations. In the labs the most common schema has been to let subjects with no history of vocal problems read a standard text at different loudness levels over varied time periods [1-3]. Some studies have used natural work as a vocal load [4, 5]. The results vary in relation to the research paradigm used, some showing a clear trend between loading and fatigue, some not. One of the most extensive examinations of the relationship between vocal load and fatigue is the study by Pekkarinen et al. [3], where the working day of a teacher was simulated. Not all the results can be unambiguously interpreted because of the great interpersonal and intrapersonal variation, but one of the parameters that was most clearly affected by the load was the F0.

Although laboratory studies have many advantages due to the possibility of controlling independent variables, the problem of generalisation remains unsolved. From every day life we all know that the voice is sensitive to different situations, to our own moods, and to our personality, and this has been verified in investigations with acoustic methods [6-10]. Undoubtedly there are features in natural contexts that influence behaviour and voice and that cannot be created in labs. However, the problems of suitable equipment and of analysis methods may in part have hindered the development of field studies. In Sweden two interesting examinations were arranged in working situations: one was carried out with electroglottography [6], the other with a voice accumulator, i.e. a small contact microphone fixed to the anterior neck [7]. Voice pitch, its range and phonation time were the main variables measured in these studies. The results were promising and they showed that of these parameters the F0 can indicate changes in the vocal load.

The aim of the present preliminary study is to improve the objective description of voice in professional voice users by developing a method for use in occupational circumstances.

METHOD

Subjects and Recording Procedure

Ten female teachers, mean age 45 years (range 33 - 53), in the junior grades of a Finnish school (schoolchildren aged 7 - 12) participated voluntarily in the study. The mean time for having been a teacher was 19 years (range 6 - 30). 28-item questionnaires charted both the teachers´ subjective appraisal of their voice problems and the background variables. All the subjects were nonsmokers and no one was undergoing voice therapy at the time of their participation. A phoniatrician examined the larynx of all except two teachers who could not find a suitable time for examination. None of the examined persons had pathology in the larynx except one whose vocal folds did not close fully, which was evaluated, however, to be a normal variation of the female laryngeal function. The questionnaires revealed that teachers experienced widely varying numbers of symptoms of vocal fatigue.

The speech samples of the teachers were collected using a battery-operated portable DAT-recorder (Sony TCD-D3) that permitted the teacher to walk freely in the classroom. The microphone was attached to a head-band and was located to one side of the mouth, 6 - 8 cm from the lips. The teachers recorded their first and last lessons of the same day. The average duration of the lessons was 35 minutes.

Data Analysis

The F0 and sound pressure level (SPL) were measured with a commercially available analogical system (the modular series by F-J Electronics, Inc.) consisting of a pre-amplifier, a F0 meter, an intensity meter and an audio frequency filter. For the F0 analysis the signal was low pass filtered at 330 Hz with the slope of 36 dB/octave and high pass filtered at 70 Hz and amplified. The logarithmic output of the meter was used in all subsequent data processing and analyses. For the calibration in the SPL analysis a sine wave of 200 Hz with the intensity of 80 dB was used. The signal was amplified (Sony PCM-F1) and input into two channels of the intensity meter with different integration times (2.5 ms and 10 ms). The signal of the shorter integration time was used for computing speech and pause times, the longer one for measuring the SPL. Altogether four signals were needed in the analysis.

After measuring the signals they were analysed by a micro-computer (Apple Macintosh Quadra 950) equipped with three extension boards (a National Instruments MIO-16-9L data acquisition board, a DSP2300 signal processor board and a DMA2800 direct memory access board). The digital speech processing functions were provided by software blocks custom-built and implemented to the LabVIEW 2 graphical programming system (National Instruments, Inc) consisting of four programs: calibration (see above), data acquisiton, editing and analysis.

In the data acquisition program the signals were digitized with the sampling rate of 5 kHz to each signal. The maximum duration of one input speech sample was 4 minutes because of the limitations of the computer memory. Samples were taken from the beginning, middle and end of the lesson. The editing program was used to exclude undesirable distorting background noise from the signal. Graphical display of the signal in a scrollable time window (each of the four data channels could be selected) and the monitoring of the sound through headphones permitted the erasure of any unwanted parts. For control purposes it was also possible to measure directly the F0 and the SPL values. The analysis program calculated mean values and standard deviations of the F0, SPL and speech and pause times. The limits employed in this study were 61 - 100 dB for the SPL and 140 - 450 Hz for the F0, values falling outside these limits being ignored. The shortest durations that were identified as speech segments were 70 ms long and as pause segments 250 ms long.

For the statistical analysis the Wilcoxon matched-pairs signed-ranks test and the Pearson correlation coefficient were used.

RESULTS AND DISCUSSION

The teachers found the practical arrangements of this study fairly unobtrusive. The tape-recorder was light to wear and easy to use, and recording did not interfere in the normal activities during the lessons. Further details about experiences and solutions to emerged problems are discussed in Rantala et al. [14].

Athough the results did not generally reach a statistically significant level, trends could be found. The overall mean voice pitch used by the teachers was 232 Hz. However, the mean values of F0 for the first and last lessons were different: it was ten hertz higher at the end of the working day (p=0.139). The difference between the beginning of the first lesson and the beginning of the last lesson was significant (p=0.059), Fig. 1. The differences between the beginning of the first lesson and the middle of the last lesson (p=0.16), and between the middle of the first lesson and the middle of the last lesson (p=0.09) were interesting though not significant.

The correlation coefficient between the F0 of the two lessons was significant (r=0.779, p=0.008): the direction of the change of the F0 was

similar, i.e. the voice pitch rose towards the end of the day for most of the teachers.

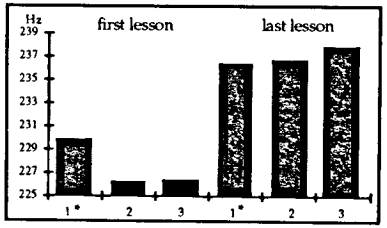


*Figure 1. The mean values of the F0 at different times of the working day in 10 teachers. Sample duration is 4 minutes. 1=beginning, 2=middle, 3=end of the lesson. *=difference significant, p=0.059*

The F0 curves of the teachers were interesting and informative from the clinical point of view. The within-teacher pitch range varied considerably, being quite minimal in some teachers, wide in others. The teachers' mean F0 values for four-minute periods varied during lessons, the lowest value being 180 Hz in one subject and the highest even as high as 293 Hz in another. Both values were considerably over the 95 % confidence interval of the mean F0, which was in this data 215 - 250 Hz. It seems that voice production in the teaching situation has great variation, and is quite different compared to that of sustained vowels produced in peaceful circumstances. Interestingly, in another related project, the mean F0 of the same subjects was found to be 186 Hz as measured from a prolonged /a/ during teaching breaks in the working week [11].

Athough there is great variation, some trends can be seen. The teachers with most difficulties in using their voice had either a high F0 for all of the analysed time in both lessons or their F0 rose towards the end of the working day. The two teachers with only a few symptoms of vocal fatigue displayed the lowest F0 values and the F0 had no sudden or large changes during the day. One teacher who subjectively felt only a few symptoms had the most deviating F0 curve of all. There were large differences in the mean F0 values for the four-minute spans, at most as much as 79 Hz. Figs. 2 a and b present the F0 curves of three teachers with the least symptoms of vocal fatigue, and three teachers with the most obvious .

In a field study there are unavoidably numerous background variables that cannot be controlled and their effects only guessed at. In laboratory settings, on the other hand, time and noise can be exploited as independent variables. In the present study the talking time of the teachers in lessons was almost half of the measured time (42 % in the first and 45 % in the last lesson) and it had no correlation to the F0. When comparing this result to other studies [7, 12] where the speaking time of a work-day has been measured, it is not surprising to find that teachers talk longer than other professions do.

The other variable that could explain the rise in the F0 is the increased loudness of voice because of background noise [13]. The mean values of the SPL of the first and last lessons varied from 78 dB to 81 dB, a correlation existing between the parameters, but not in the two last measured four-minute periods when the F0 was at its highest. This phenomenon probably results more from vocal fatigue than from increased loudness.

Although the results of this preliminary study of field-study methods are promising, they must be considered with caution because of the small sample and the inherent limitations that belong to all field studies. Many background variables (e.g. the noise of the classroom, the teacher's teaching and speaking style, talking time) could not be controlled and regulated systematically. On the other hand some variables are present in all studies of human behaviour (e.g. personality, experienced stress, physical conditon) and they can be controlled neither in field studies nor in laboratory conditions.

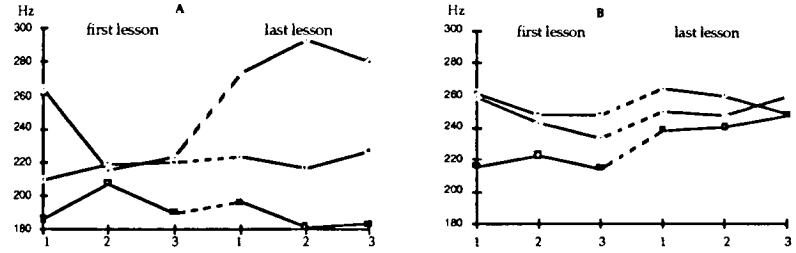


Figure 2. Mean F0 values for six teachers. a) Three teachers with the least symptoms of vocal fatigue. b) three teachers with the most obvious symptoms of vocal fatigue. 1= beginning, 2=middle, 3=end of the lesson.

Further studies are planned in order to increase the number of subjects and hence improve the reliability of the results. It appears that the method is suitable for clinical practise, for specifying diagnoses more closely, focusing the aims of voice therapy and allowing the better following-up of rehabilitation.

REFERENCES

[1] Reimers Neils, L., Yairi, E. (1987), Effects of speaking in noise and vocal fatigue and vocal recovery, *Folia Phoniatr*, vol. 39, pp. 104-112.

[2] Pausewang Gelfer, M., Andrews, M.L., Schmidt, C.,P. (1990), Effects of prolonged loud reading on selected measures of vocal function in trained and untrained singers, *J Voice*, vol. 5, pp. 158-167.

[3] Pekkarinen, E., Alku, P., Lauri, E-R., Nykyri, E., Sihvo, M., Toivonen, P., Vilkman, E. (1993), *Separate and joint effects of vocal load in working environment on voice production in professional voice users*, in Finnish, published by support by the fund of occupational safety and health, project 91010, Helsinki.

[4] Löfquist, A., Mandersson, B. (1987), Long-time average spectrum of speech and voice analysis, *Folia Phoniatr*, vol. 39, pp. 221-229.

[5] Novak, A., Dlouha, O., Capkova, B., Vohradnik, M. (1991), Voice fatigue after theater performance in actors, *Folia Phoniatr*, vol. 43, pp. 74-78.

[6] Kitzing, P. (1979), *Glottographic frequency analysis*, in Swedish, Dissertation, Lund University.

[7] Ohlsson, A-C. (1988), *Voice and work environment: towards an ecology of vocal behaviour.*, Dissertation, Gothenburg University.

[8] Novak, A., Vokral, J. (1993), Emotions in the sight of long-time averaged spectrum and three-dimensional analysis of periodicity, *Folia Phoniatr*, vol. 45, pp. 198-203.

[9] Brenner, M., Doherty, E.T., Shipp, T. (1994), Speech measures indicating workload demand, *Aviat. Space Environ. Med.* vol. 65, pp. 21-26.

[10] Freidl, W., Friedrich, G., Egger, J. (1990), Persönlichkeit und Stressbearbeitung bei Patienten mit funktioneller Dysphonie, *Folia Phoniatr*, vol. 42, pp. 144-149.

[11] Tumanoff, H. *Perturbation as an indicator of vocal fatigue in teachers*, in Finnish, Master's thesis, Oulu University, under preparation.

12] Watanabe, H., Shin, T., Oda, M., Fukaura, J., Komiyama, S. (1987), Measurement of total actual speaking time in a patient with spastic dysphonia, *Folia Phoniatr*, vol. 39, pp. 65-70.

[13] Schultz-Coulon, H-J. (1980), Zur routinemässigen Messung der stimmlichen Reaktion im Lärm, *Sprache-Stimme-Gehör*, vol. 4, pp. 28-34.

[14] Rantala, L., Haataja, K., Vilkman, E., Körkkö, P. (1994), Practical arrangements and methods in the field examination and speaking style analysis of professional voice users, *Scand J Log Phon*, vol. 19, pp. 43-54.

TREATING CHILDREN WITH PERSISTENT ARTICULATORY PROBLEMS USING INDIVIDUALLY DESIGNED PALATAL PLATES AD MODUM CASTILLO-MORALES

Anita McAllister, speech pathologist & Martha Björnström, doctor of dental surgery
Dept. of Speech Pathology, Danderyd Hospital, S-182 88 Danderyd, Sweden.
Ph +46 8 655 58 11, FAX +46 8 622 58 57.

Abstract
In an attempt to evaluate treatment with palatal plates for children with primarily articulatory dysfunction 20 children were treated.. Each child got at least one follow-up appointment after a period of two to three months.
Fourteen children had improved articulation at their first check-up after 3 months.
The palatal plate intensified the exercise program and led to more rapid results.

INTRODUCTION

Persistent articulatory problems have long been a challenge to most speech pathologists working with children. Despite substantial effort on both parts the problem may still remain. During the last two decades this group of patients have received increasing attention and their
special types of difficulties has been studied by several researchers, for example [1, 2, 3]. Different therapy techniques have also been described usually involving simple oral motor exercises with increasing complexity as the treatment proceeds, [4] The term verbal apraxia was proposed by Morley, [5]. A generally accepted definition of the problem is that it is a neurological disorder affecting speech production without any overt motor or sensory paralysis

In our clinic an intraoral treatment with a palatal plate or a vestibular brace has been one possible method of treating children with persistent articulatory problems. The method was originally established in order to improve the rehabilitation of patients with oral motor problems due to trauma or stroke. The purpose of the plate was initially to induce and facilitate swallowing and the method has been described by dental surgeon Selley, [6]. The concept of intraoral treatment devices with a more wide spread use has also been proposed by Dr. R Castillo-Morales, [7] this has also been described by Hoyer and colleagues, [8, 9].
The purpose of the present investigation was to evaluate the treatment of palatal plates for children with articulatory problems.

METHOD

Twenty children participated in the investigation, 6 girls and 14 boys. The mean age when starting their treatment with the palatal plate was 8,4 years with a standard deviation of 3,58 years ranging from 5,0 - 16,0. All children had an history of previous treatment by speech pathologists or therapists based on various methods. A thorough evaluation was made including both sensory and motor tasks. The sensory test included tasks such as identifying two tactile stimuli on the tongue with varying inter stimulus distance. The motor task was mainly imitation of simple oral movements, like blowing and sucking. Articulatory tasks involved single sound production and monosyllabic words (CV) to establish possible difficulties with certain articulatory targets. Dynamics were tested with sequences of sounds like /lalalala/ or /tititiii/. The ability to change place of articulation was tested with a sequence of /patakapatakapataka/. Auditory discrimination was tested in order to establish if this disability was part of the problem complex. A speech sample was collected. This involved an informal interview with the patient and for those with reading skills a short text. Frequently the oral motor problem is combined with other difficulties such as comprehension difficulties, general sequencing difficulties, problems with eating or sucking and voice problems. Children with oral dyspraxia are not aware of the exact position of their articulators. Often they have to check the result with their hands or in a mirror. We also found a tendency for dyspractic children to overdo the exercises involved in the test. None of these children had any substantial problems involving mimic muscles or auditory discrimination although dynamics - rapid movements of the tongue - was often part of their problem complex. Eight children had problems with the sensory feedback from the tip of the tongue as tested with two stimulator at different distances.
Prior to the treatment with a palatal plate 19 children participated in a regular treatment program for an average of 5 months. This program involves oral exercises with lips, tongue and soft palate, vibration stimulation as well as oral exercise with resistance, simple articulatory tasks and exercises involving the mimic muscles.

Five children had problems mainly concerning /r/. The aim of their plate was to reinforce the stimuli of the articulatory place for the tongue when producing /r/. Ten children had problems mainly involving the upper lip and the tip of the tongue. A palatal plate was made to stimulate the tongue and normalize the mobility of the upper lip.
Three children dentalized the velar sounds /k, g/ and two children had a lateral /s/-articulation. One of these children also had a palatalized articulation of /l/.
The palatal plates were individually designed to train each child's specific articulatory deficits. The palatal casts were made in an upright position to avoid breathing difficulties and the possible feelings of panic. Each child got at least one follow-up after a period of 2,5 - 3 months. Improvements led to alterations of the palatal plates or a termination of treatment.

RESULTS

The evaluation of treatment results was made by the authors in collaboration with the parents and patients at each follow-up. It was documented with video or photography.
In the first group two children succeeded in establishing a trilled [r]. One mother reported that the [r] was established after using the palatal plate for a period of only 10 days. The other 3 had all succeeded in producing an acceptable [ʒ] at the time of the first or second control. This is a sound that is accepted as /r/ in standard Swedish.

N	Norm	Improve	No change
/r/ 5	2	[ʒ] 3	
/t,d,n/ 10	3	7	
/k,g/ 3		2	1
/l,s/ 2		1	1
Tot	5	13	2

Table 1. Results at the first and/or second check-up after 2,5 - 3 and 6 months of treatment respectively. None of the problems got worse.

Three patients in the second group had normalized their dental articulation and treatment have thus terminated. Seven children had improved articulation meaning that the velar articulation was more palatalized or retroflexed
In the third group 2 children improved their articulation but no one was normalized. One child had not changed the place of articulation and still dentalized all velar stops.
The two children in the fourth group had problems with lateral /s/ and the l-sound. The child with difficulties articulating l improved his articulation. However he did not use the edge of the tongue but rather had the tongue a little bit interdental when articulating l, thus it could not be considered to be normalized. The lateral s articulation still remained at the time of the first check-up for both children.
All patients in this study could adapt to the regular use of an intraoral device.

Discussion

The palatal plates are used to strengthen the intraoral sensory feedback when that is necessary. It gives the child a clear target to aim at. The use of a palatal plate does not put the same demands on the parent as a regular exercise program. The patient and the plate do the work by themselves. This is invaluable in patients with long treatment periods For the patient this means a possibility to decrease the time spent with the speech pathologist in the clinic. this can be invaluable for this group of patients who all have had exstensive speech therapy.
Many patients with oral motor problems may also have eating and chewing difficulties. In our group of patients this could not be noticed. However in their case history chewing and eating difficulties had take place at some point for 6 patients.
In our project all patients could adapt to the use of an intraoral device. This is in good agreement with previous studies that report good results with very young children often not a year of age [7, 8, 9]. Needless to say the aim of these plates have not primarily been articulation but rather eating, sucking and reduced drooling for example.
A positive side effect of the plate for some children hypersensitive in the back of the moth has been facilitated tooth brushing of the back molars. We have also noted positive psychological effects when the patients themselves note the improved articulation.

Conclusion

- Children with articulatory problems treated with a palatal plate should not have difficulties with auditory discrimination.
- A palatal plate can help establish the place of articulation for different phonemes by strengthening the stimulation of the articulatory target.
- It was easier to adapt to a regular use of the palatal plate than to establish regular habits with articulation exercises.
- The palatal plates seemed to intensify treatment and lead to more rapid results.

References

[1] Ferry PC, Hall SM, Hicks JL. (1975)."Dilapidated" Speech: Developmental Verbal Dyspraxia.*Develop Med Child Neurol, 17, 749-756.*
[2] Prichard CL, Tekieli ME, Kozup JM. (1979) Developmental Apraxia: Diagnostic Considerations. *J of Com Dis, 12, 337-348.*
[3] Amorosa H, von Benda U, Wagner E. (1990). Voice Problems in Children with Unintelligible Speech as Indicators of Deficits in Fine Motor Coordination. *Folia Phon,* vol 42, pp64-70.
[4] Bashir AS, Graham Jones F, Bostwick RY. (1984). A touch-cue method of therapy for developmental verbal apraxia. Seminars in Speech and language 5, 2, 127-137.
[5] Morley MF: (1965). The Development and Disorders of the Speech in Childhood. 2nd ed, Williams and Wilkins, Baltimore.
[6] Selley W G. (1985). Swallowing difficulties in stroke patients: a new treatment. *Age and Ageing 14,* 361-365.
[7] Castillo-Morales R. (1989). *Orofaciale regulationsterapie.* Pflaum Verlag, München.
[8] Hoyer H, Limbrock G J. (1990). Orofacial regulation therapy in children with down syndrome, using the methods and appliances of Castillo-Morales. *J of dentistry for children, Nov. - dec.*
[9] Limbrock G J, Hoyer H, Scheying H. (1990). Drooling chewing and swallowing dysfunction in children with cerebral palsy: Treatment according to Castillo-Morales. *J of dentistry for children, Nov.- dec.*

JITTER AND SHIMMER IN VOCAL FOLD NODULES, POLYPS AND EDEMAS BEFORE AND AFTER PHONOSURGERY

*A. Nieto; A.Vegas; F.Gamboa; J. Montojo; I. Cobeta and P. Kitzing**
Hospital Universitario "Príncipe de Asturias", Universidad de Alcalá. Madrid .Spain.
**ENT Department, Malmö General University Hospital, Malmö, Sweden*

ABSTRACT

Perturbation analysis in patientes suffering from nodules, polyps and Reinke's edemas are studied by means of acoustic (MC), electroglottographic (EGG, Lx) and flow glottographic (FGG) waves before and after microlaryngoscopic phonosurgery.

INTRODUCTION

According to the myoelastic theory, pressure below the vocal cords increases till the glottis opens setting the vocal folds in vibration. In the light of the source-filter theory of speech production, larynx is the source with the vocal folds chopping the column of exhaled air. In this way the "buzz" generated in the glottis, becomes audible as a vocal sound, outside the lips, by the action of the vocal tract, that behaves as a filter.

Vibratory pattern of the vocal folds is not always regular. **Perturbation** is the term applied to these deviations from regularity. Multiple indices for measuring perturbation have been developed. Most of them represent some sort of average of the difference between the periods (jitter) or amplitudes (shimmer) of succesive vocal cycles [1].

Since the first attempts to create objective measures for perturbation analysis by Von Leden , research on this matter has generated numerous papers, a comprehensive overview of them can be found in Laver et al [2].

Several factors have been suggested as posible contributors to irregularity in vocal fold vibration [3]: 1) unsteadiness in muscle contraction in the laryngeal and respiratory system; 2) turbulence in glottal air stream; 3) instability in the jet emerging from the glottis; 4) asymetry in the mechanical or geometrical properties of the two vocal folds; 5) nonlinearity in the mechanical properties of the vocal fold tissues; 6) changes in coupling between the vocal fold and the vocal tract and 7) mucus riding on the surface of the vocal folds.

If perturbations are present in normal phonation, it stands to reason that jitter and shimmer should be increased in the presence of vocal fold pathology. So perturbation could be employed to detect vocal fold pathology and to evaluate the resulting disordered voice.

Nodules, polyps and Reinke's edemas are common findings of ENT practice. Management of this vocal abuse pathology (VAP) includes voice therapy and surgery.

Phonosurgery (PS) refers to surgical techniques, designed to improove or restore the voice, based in the three layer structure of the vocal folds, .

The first purpose of the study was to establish the clinical usefullness of perturbation measures to follow patients that underwent phonosurgery. The second is to investigate the influence of the type of voice signal employed for perturbation calculus: microphonic (MC), laryngographic (Lx), inverse filtered (FGG). We studied RAP (relative average perturbation factor) [4] for frequency perturbation and shimmer in dB for amplitude perturbation [5].

METHODS

1. **Signal acquisition hardware**. Signals from a microphone (MC) (Shure Prologue), a Fourcin's Laryngograph (Lx) (Kay Elemetrics) and an inverse filtering system (FGG) provided with a Rothenberg mask (Glottal Enterprises) and a filtering system designed by the Tech. Dep. of Malmö General Hospital; were digitized two by two (MC, Lx) (FGG, Lx) on a CSL 4300 (Kay Elemetrics) data acqisition system provided with a 16 bits A/D converter, with a sampling rate 51200 Hz. using a 486, 8 Mb RAM PC.

2. **Test procedures**. All tests were carried out in three different moments: before surgery, 2 weeks and 1 month after surgery. a) Subjects. 52 patients (32 polyps, 11 Reinke's edemas and 9 nodules) were examined using laryngostroboscopy. Age ranged from 65 to 15, mean 40. Sex was 22 (M): 30(F).
b) Recording. Recordings made in a sound treated booth were stored in a rewritable optical disck 1.3 GB.
c) Voice tasks. The first task was to produce 2s of a sustained /a/ with Rothenberg mask and laryngographic electrodes in position, at a confortable pitch and loudness. Both signals (FGG,Lx) were simultaneously recorded. The second task was to produce 2s of a sustained /a/ at 5 cm mouth-to-microphone distance with laryngographic electrodes in position, at a confortable pitch and loudness, a simultaneous recording of both signals (MC, Lx) was made. This second task was performed twice (acoustic analysis II and III).
d) Analysis. analysis was based on sustained /a/ because formant configuration in this vowel was more suitable for inverse filtering proccedure.

Relative average perturbation (RAP) was used for jitter measures, and shimmer in dB for calculus of amplitude perturbation. Initial and terminal portion of phonation were excluded and only 2s of the remaining stable portion were used for analysis.

3. **Statistic analysis**. The first part of the analysis consists of descriptive statistics to obtain location measures (mean, S.D). The second part include U-Man-Whitney rank sum tests for ANOVA of unpaired data, and T-Wilcoxon tests for ANOVA of paired data.

RESULTS AND DISCUSSION

Table 1. Results: (MC.) acoustic wave; (Lx) Lx wave and (FGG) flow glottogram . Jitter(J) in %, shimmer (S) in dB. ; Pre- surgery(p) and 1month - Post-Surgery(m).

	Mean S.D.	Mean S.D.	
MCJp	2.075 1.170	1.904 1.791	MCJp
MCJm	0.460 0.543	0.510 0.301	MCJm
MCSp	0.575 0.390	0.566 0.307	MCSp
MCSm	0.257 0.092	0.354 0.107	MCSm
LxJp	0.892 0.840	1.320 1.389	LxJp
LxJm	0.364 0.193	0.418 0.301	LxJm
LxSp	0.454 0.453	0.536 0.392	LxSp
LxSm	0.217 0.135	0.208 0.168	LxSm
FGJp	2.046 2.167	1.904 1.791	FGJp
FGJm	0.768 0.428	0.510 0.301	FGJm
FGSp	0.775 0.967	0.566 0.307	FGSp
FGSm	0.255 0.137	0.354 0.107	FGSm
	POLYPS n=32	EDEMAS n=11	p < 0.05

Increased perturbation in voice may result from unsuitable patterns of vocal fold vibration, induced by the presence of pathology. Attempts of objective evaluation of voice in patients were restricted to medical research voice labs. Recent computer development has made this attempt in the clinic reasonable.

We evaluated the effect of nodules, polyps and edemas in the capacity for regular vibrations of the vocal folds. A nonparametric ANOVA test (T-Wilcoxon) for paired data was used to establish the influence of choosing a determined production of the vowel /a/ of the different possible trials. There were no significant differences in jitter and shimmer values, supposed the same type of voice signal was employed (MC, Lx or FGG).

The same test showed significant differences ($p<0.05$) for jitter and shimmer depending on the type of wave

used for perturbation calculus. Our results are opposite to those found in other articles [6] where these differences proved erratic. According to our results FGG-jitter values were always higher than Lx-Jitter ($p<0.05$). Differences could be owed to the manual process for filtering the flow glottographic signal, paricularly before PS when the register was more irregular and a correct filtration specially difficult to achieve. Despite this we think differences are basically due to the distinct nature of the phenomena represented by both waves.

Patients with polyps showed before PS superior values of shimmer computed on FGG basis than Lx basis. But two weeks and one month after PS Lx-shimmer was superior to FGG-shimmer. A possible explanation for this is that once the lesion is excised, the small volumes of air liberated with each vocal cycle, represented by the peaks of the FGG, should be more uniform, as the glottal closure inproves and mechanical balance of vocal folds is restored. In the case of Lx wave peaks could be contaminated with artefacts. These artefacts, even present before PS, are probably masked by the superior grade of variation due to the lesion presence.

Patients with Reinke's edema showed before PS higher values for FGG-shimmer than Lx-shimmer. One month after PS again FGG-shimmer is greater than Lx-shimmer. The latest statements seem in contradiction with the reasons argued in the preceding discussion in patients with polyps and have difficult explanation. In their interpretation the different performance of polyps and edemas showed by stroboscopy should play a role . One month after PS, stroboscopy proved that: free margens of the vocal folds were more irregular, glottic closure was more incomplete and the presence of inflamatory signs were more evident in the case of edemas than in the case of polyps.

MC-Jitter was superior to Lx-Jitter independently of the type of lesion present in the vocal fold, before and after PS ($p<0.05$). A posible reason for this is that the acoustic wave is more complex than the Lx wave. Presumably multiple deflections of the MC-wave involves a major difficulty for pitch extraction algorithm, to identify the peak on which to calculate the period than in the case of the Lx wave, with a single deflection. Superiority of Lx wave above acoustic wave for pitch extraction has been mentioned by other authors [7].

A U-Mann-Whitney test for unpaired data showed no significant differences between different pathologies. Our findings are in agreement with previous studies [6,8,9] supporting that perturbation cannot be used to distinguish among several pathologies.

We have not found differences between men and women in perturbation values. Our results differ from others that find greater jitter values in women [10,11]. We agree with others who find relations between amount of jitter and gender somewhat equivocal [12].

Data presented show an appreciable difference between perturbations found before PS and jitter and shimmer found in normal speakers, when the same algorithms were used, either in the case of shimmer [5,13,] or in the case of jitter [6,13] . When the lesion is present , mechanics of the two folds are different and more irregular vibrations are the result. Added to this is the incomplete closure of the glottis due to the presence of the lesion.

Data presented support that jitter and shimmer values were clearly inferior after PS. ANOVA tests for paired data (T-Wilcoxon) proved this significant ($p<0.05$). This downward shift could be expected and if we take into account that pathologic voices have an increased amount of perturbations , it's obvious from our data that voice quality of our patients clearly improved. So primary

intention of phonosurgery was achieved. The fact that published normal values of jitter and shimmer are similar and even superior to the values obtained in our study, provide more evidence for definite voice quality inprovement in our patients.

For a correct interpretation of our results it should be noticed certain differences of our study with refered works. For analysis we employed a sustained /a/ while others used the vowel /i/ [6,15] and with older subjects than our patients (mean age 40).

CONCLUSIONS

1.- Jitter and shimmer are increased in nodules, polyps and edemas no matter the type of wave used for the analysis (MC, Lx or FGG).

2.-Perturbation analysis does not make differential diagnosis among different pathologies.

3.- Perturbation values do not depend on the trial chosen for analysis (supposed the same evolution moment of the study is compared), but depends on the type of wave used.

4.- Perturbation analysis is a usefull method to evaluate results in phonosurgery.

REFERENCES

[1] BAKEN, R. J. (1987), "Clinical measurements of speech and voice" Boston: Little Brown.

[2] LAVER, J.; HILLER, S.; MACKENZIE, J. (1992), "Acoustic waveform perturbation and voice disorders" J. Voice vol. 6(2), pp. 115-126.

[3] TITZE, I.R.; BAKEN, R.J.; HERZEL, H. (1993), "Evidence of chaos in vocal fold vibration" in Titze,I.R. "Vocal fold phisiology" San Diego: Singular Publishing Group pp. 143-188.

[4] KOIKE, Y. (1973), "Applications of some acoustic measures for the evaluation of laryngeal dysfunction" Stud Phonol vol. 7, pp 17-23.

[5] HORII, Y. (1980), "Vocal shimmer in sustained phonation" J. Speech Hear. Res. vol. 23, pp 202-209.

[6] LABLANCE, G; MAVES, M.D.; SCIALFA, TH.M.; EITNIER, C.M.; STECKOL, K.F. (1992), "Comparison of electroglottographic and acoustic analysis of pitch perturbation" Otolaryngol. Head Neck Surg. vol. 107, pp. 617-621.

[7] ASKENFELT, A; GAUFFIN, J; SUNDBERG, J; KITZING, P (1980), "A comparison of microphone and electroglottograph for the measurement of vocal fundamental frequency" J. Speech Hear. Res.vol. 23, pp. 258-273.

[8] LUDLOW, C.L.; BASSICH, C.J.; CONNOR, N.P.; COULTER, D.C.; LEE, Y.J. (1987), "The validity of using phonatory jitter and shimmer to detect laryngeal pathology" In Baer, T.; Sasaki, C.; Harris, K.S. eds. "Laryngeal function in phonation and respiration" Boston: Little Brown, pp 492-508.

[9] FEIJOO, S.; HERNANDEZ, C. (1990) "Short term stability measures for the evaluation of vocal quality" J. Speech Hear. Res. vol. 33, pp 324-334.

[10] IWATA, S.;VON LEDEN, H. (1970), "Pitch perturbation in normal and pathological voices" Folia Phoniatr. vol. 22, pp. 413-424.

[11] JAFARI,M.; TILL,J.A.; TRUESDELL , L.F.; LAW-TILL,C.B. (1993), "Time shift trial and gender effects on vocal perturbation measures" J. voice vol. 7(4), pp. 326-336.

[12] SUSSMAN, J.E.; SAPIENZA, C.H. (1994), "Articulatory, developmental and gender effects on measures of fundamental frequency and jitter" J. voice vol. 8(2), pp. 145-156.

[13] SORENSEN, D.; HORII, Y. (1983), "Frequency and amplitude perturbation in the voice of female speakers" Journal of Communication Disorders vol. 16, pp. 57-61.

[14] CASPAR,J.C. (1983), "Frequency perturbation in normal speakers: A descriptive and metodological study" Doctoral Disertation. Syracuse University. Syrakuse. N Y.

THE EFFECT OF RADIOTHERAPY ON VARIOUS ACOUSTICAL, CLINICAL AND PERCEPTUAL PITCH MEASURES

Irma M. Verdonck-de Leeuw & Florien J. Koopmans-van Beinum
Institute of Phonetic Sciences/IFOTT, Amsterdam, The Netherlands

ABSTRACT

Speech samples of patients with early glottic carcinoma and of control speakers, are analysed for acoustical pitch and EGG, as well as by means of perceptual pitch evaluation by trained and untrained raters. The results of pitch ratings by trained listeners, EGG, and acoustical pitch correlate strongly and show the tendency that voices before radiotherapy have a higher pitch than 6 months and 2 years after radiation. Voices longer than 3.5 years after radiation tend to become higher again.

INTRODUCTION

Within the scope of a co-operative study between the Netherlands Cancer Institute (Antoni van Leeuwenhoek Hospital), the Academic Hospital of the Free University of Amsterdam, and the Institute of Phonetic Sciences of the University of Amsterdam, research is carried out on the effect of radiotherapy on voice quality. The aim of this study is to obtain parameters that can describe voice quality of patients with early glottic cancer before and after radiotherapy and of normal speakers. Voice quality can be described by several perceptual, clinical, as well as acoustical methods. In this presentation we will focus on various pitch measures. Data on acoustical pitch and EGG pitch are taken into account as 'objective' pitch measures; the results will be compared with perceptual evaluations of trained and untrained raters. The trained raters are used to provide an analytic description of voice quality. The role of the untrained raters is to find out how 'ordinary' people evaluate voice quality.

In a later stage of the study the results will be compared with other perceptual parameters of voice quality (evaluations on semantical scales, such as breathiness, harshness, creakiness), with other clinical methods (phonetogram, phonation quotient, and evaluation of stroboscopic recordings of vocal fold vibration), and with other acoustical analyses (LTAS, SNR, perturbation).

Before radiotherapy, the actual tumour can cause changes of the vocal folds such as mass change, stiffness change, and asymmetry. Little is known about voice quality after radiotherapy. Some studies report voice improvement to a normal or near-normal level, 6-12 months after radiotherapy, for about 70% of the patients [1,2,3]. Other studies report abnormal post-radiation voices [4,5]. Pitch may be one of the parameters that can be influenced by the presence of a tumour or by the effect of radiotherapy on the vocal fold tissue, such as late oedema, necrosis etc. Furthermore, pitch measurements are important cues for other acoustical and perceptual measurements, such as spectral noise [6,7,8], breathiness, and tension [8,9].

METHOD

Speakers/recordings

Patients with early glottic cancer (T1N0M0) are treated with radiotherapy (60 Gy in 30 fractions, or 66 Gy in 33 fractions). Voice samples of the same 10 patients have been recorded before radiation, as well as 6 months, and 2 years after radiation. Recordings are also made of 3 other groups of 10 patients each, before radiation, 6 months, and 2 years after radiation, and of 20 patients longer than 3.5 years after radiation. Finally, recordings are made of 20 control speakers without any known vocal defects (figure 1). The matching between patients and control speakers took into account sex (all male), age (47-81), as well as smoking habits. The speakers read aloud a text for about 5 minutes. All the material was recorded using a Casio DAT-recorder and a Philips N8214 microphone. Fragments (ca 30 sec.) of all texts were digitised by means of the Sound editor of an Iris Indigo R4000, sample frequency 48 kHz, 16 bit resolution. These samples were copied to cassette tapes in two random speaker orders.

	PRE	P1	P2	P3	
Long.	10 ->	10 ->	10		
Mix.	10	10	10	20	
Control					20

Figure 1. Illustration of the various speaker groups. Longitudinal group before radiation (PRE), 6 months after (P1), and 2 years after (P2). Mixed groups PRE, P1, P2 and P3 (longer than 3.5 years after radiation). Control group.

Raters/rating procedure

The untrained raters in this experiment are 20 university students (6 male, 12 female), without any experience for this listening task. They were paid for their participation. The raters received written instructions. First they heard examples of 10 different voices in order to get a reference frame. After the examples 110 fragments of read aloud text were presented (10 training fragments and 100 fragments of speakers as indicated in figure 1). The raters judged voice quality on 14 voice quality scales. The tapes were presented binaurally via a cassette recorder and headphones. The raters listened to the tapes in a quiet room, individually. On the average, the whole rating procedure (instructions + rating) took about 1 1/2 hours.

The trained raters are 3 female (socio-) phonetic researchers; 2 had followed a training course on the Voice Profile by John Laver. They rated the voices on 8 scales independently from each other.

The semantic scales consist of various voice quality scales that have been used in previous experiments [9]. In this paper we limit ourselves to the 'pitch' scales *low-high* and *deep-shrill* (7-points) used by the untrained raters, and *low-high* and *sonor-shrill* (13-points) used by the trained listeners.

Acoustical Pitch

Average pitch values of the read aloud text are determined by means of the program "Praat" [10]. The acoustic pitch period of a sound is determined by the position of the maximum of the autocorrelation function of the sound. This procedure is extensively described by Boersma [10]. In order to select only voiced candidates for pitch detection the Voicing Threshold is set to 0.5 and the Silence Threshold to 0.05. All other parameters are kept default.

Clinical fundamental frequency

By means of an electroglottograph (Stopler Teltec GFA06) the average fundamental frequency is measured for the read aloud text. The speakers read aloud the same text for about 5 minutes while 1000 voiced samples were analysed and averaged.

RESULTS

The reliability coefficient Ru is used as a measure of the reliability of the means of the ratings by a panel of raters (between 0 and 1). Ru= (MSspeakers - MSraters) / MSspeakers. The results for the 3 trained raters are Ru = .80 for *low-high* and Ru = .63 for *shrill-sonor*; for the 20 untrained raters Ru = .93 for *low-high* and Ru = .94 for *deep-shrill*. The differences between trained and untrained raters lies in the low MSspeakers (=true variance) by the trained raters.

Agreement is determined by Kendall's W (between 0 and 1). For the trained raters W = 0.68 for *low-high* and W= 0.64 for *sonor-shrill*; for the untrained raters W = 0.35 for *high-low* and W= 0.30 for *deep-shrill*.

Means of the ratings per speaker per scale for the trained raters, and for the untrained raters are put into a datafile, together with the acoustical pitch data, as well as the EGG data (figure 2). Pearson correlations for the various means are given in table 1.

	2	3	4	5	6
1	.89	.46	.11	.45	.42
2		.55	.08	.57	.50
3			.24	.72	.69
4				.30	.16
5					.75

Table 1. Pearson correlations for the 6 pitch measures. 1.'low-high' untrained, 2. 'deep-shrill' untrained, 3. 'low-high' trained, 4. 'sonor-shrill' trained, 5. acoustical pitch, 6. EGG

The results of the individual data given in figure 2 show that the perceptual evaluations by trained and untrained raters do not differentiate between the speaker groups. Over all speakers, the trained raters range from 7.0 to 7.8 for *low-high* and from 6.5 to 7.4 for *sonor-shrill*. The untrained raters range from 2.7 to 3.3 for

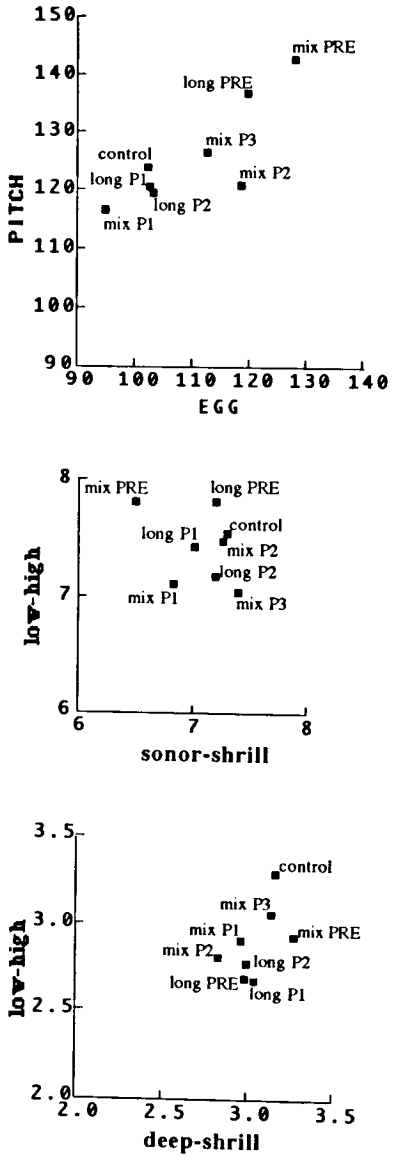


Figure 2. Means of acoustical pitch by EGG (top), and 'low-high' by 'sonor-shrill by trained raters (middle), and 'low-high' by 'deep-shrill' by untrained raters (bottom) for Longitudinal groups before radiation (PRE), 6 months after (P1), and 2 years after (P2), and Mixed groups PRE, P1, P2, and P3 (longer than 3.5 years after radiation), and control group.

low-high and from 2.8 to 3.3 for *deep-shrill*. Notice the deviating scales in figure 2: 6-8 for the trained raters (13 points scale), and 2-3.5 for the untrained raters (7-points scale).

It is obvious that the raters didn't hear specific differences between the speaker groups. This was already indicated by the reliability and agreement results: the trained raters having a high agreement score but a low reliability coefficient, due to a low MSspeakers.

The results for the acoustical pitch and EGG data do show (though statistically not significant) differences between the speaker groups. To find out wether a combination of parameters will give better insight in the effect of radiation on pitch a factor analysis is carried out.

A Principal Component Analysis is used to decompose the correlation matrix into (varimax rotated) factors (PCA). For determining the number of factors, the criterion 'eigenvalue > 1' is applied. With this criterion the PCA produced 2 factors, together explaining 80% of the total variance. On the basis of the factor loadings (> .6) the 2 factors are mainly determined by acoustical pitch, EGG, and *low-high* by trained raters (factor 1), and *low-high* and *deep-shrill* by untrained raters (factor 2). Factor scores are presented here for the first factor as it explains most of the variance (49%). The scores give the position for each speaker on each factor (figure 3). Results show that a tendency can be seen (though statistically not significant) for 'pitch' as a combination of acoustical pitch, EGG and *low-high* evaluations by trained raters: patients before radiotherapy have a very high 'pitched' voice as compared to patients 6 months and 2 years after radiation. This counts for both patient groups (longitudinal and mixed). The voices of patients longer than 3.5 years after radiation tend to become higher again, whereas the control speakers have the lowest voices.

The tendency that patients before radiation have very high pitched voices may be due to mechanical effects of the tumour on the vocal folds. Another explanation may be an increased tension of the vocal folds by the patient in order to compensate for his voice loss. Also, little is known about the effect of microlaryngeal

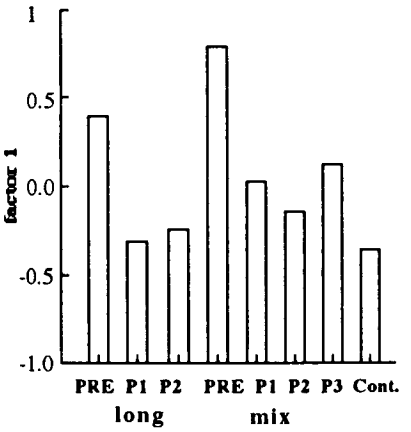


Figure 3. Means of Factor 1 for Longitudinal groups before radiation (PRE), 6 months after (P1), and 2 years after (P2), and Mixed groups PRE, P1, P2, and P3 (longer than 3.5 years after radiation), and control group.

surgery that most of the patients have undergone before radiation.

After radiotherapy the voices seem to be low pitched. Conflicting results have been found in previous research [4,9], though none of the results were significant. On the long term, the effect of radiotherapy seems to be that voices tend to become 'normal' again (2 years after radiation) but become high pitched later on. This may be due to the irradiation of the normal tissues that can result in late oedema, altering the vibratory cycle by mass and stiffness changes of the vocal folds.

Although the results in this experiment do not differentiate significantly between the speaker groups, the correlations between the acoustical pitch analysis, and the EGG data, and the pitch evaluations of the trained raters for the read aloud text are clear. The expectation that what one can hear should also be measurable, becomes true in this experiment, at least for the trained raters.

The evaluations by the untrained raters do not correlate strongly with the other analyses; still they do have an important role in this study. The purpose is to find out how 'ordinary' people (i.e. not voice reseachers/pathologists) evaluate voices of patients. Therefore in future research, the evaluations of the patients themselves and their partners are taken into account as well.

The expectation is that raters, trained as well untrained, can differentiate between speaker groups on other aspects than pitch evaluations. Obviously, pitch is not a parameter that represents strongly the effect of radiotherapy, since none of the various pitch measures discriminates clearly between the speaker groups.

REFERENCES

[1] Harrison, L.B, Solomon, B., Miller, S., Fass, D.E., Armstrong, J. & Sessions, R.B. (1990). "Prospective computer-assisted voice analysis for patients with early stage glottic cancer: a preliminary report of the functional result of laryngeal irradiation", *Int. J. Radiation Oncology Biol. Phys.* 19, 123-127.

[2] Miller, S., Harrison, L.B., Solomon, B. & Sessions, R.B. (1990). "Vocal changes in patients undergoing radiation therapy for glottic carcinoma", *Laryngoscope* 100, 603-606.

[3] Lehman, J.J., Bless, D.M. & Brandenburg, J.H. (1988). "An objective assessment of voice production after radiation therapy for stage 1 squamous cell carcinoma of the glottis", *Otolaryngol Head Neck Surg* 98, 121-129.

[4] Colton R. H., Sagerman, R.H., Chung, C.T., Yu, Y.W. & Reed, G.F. (1978). "Voice change after radiotherapy", *Radiology* 127, 821-824.

[5] Emanuel, F.W. & Smith, W.F. (1974). "Pitch effects on vowel roughness and spectral noise", *Journal of Phonetics* 2, 247-253.

[6] Wolfe, V., Cornell, R. & Palmer, C. (1991). "Acoustic correlates of pathologic voice types", *Journal of Speech and Hearing Research*, 34, 509-516.

[7] Verdonck-de Leeuw, I.M. & Boersma, P. (to appear). "The effect of radiotherapy measured by means of Harmonicity".

[8] Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J. & Wendin, L. (1980). "Perceptual and acoustic correlates of abnormal voice qualities". *Acta Otolayrngologica* 90, 442-451.

[9] Leeuw de, I.M. (1990). "The relation between perceptual and clinical parameters of voice quality of patients with early glottic cancer before and after radiotherapy and of normal speakers", *Proceedings of the Institute of Phonetic Sciences Amsterdam* 14, 27-38.

[10] Boersma, P. (1993). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", *Proceedings of the Institute of Phonetic Sciences Amsterdam* 17, 97-110.

[11] Hoyt, D.J., Lettinga, J.W., Leopold, K.A. & Fisher, A.R. (1992). "The effect of head and neck radiation therapy on voice quality", *Laryngoscope* 102, 477-480.

PERCEIVING AND PROCESSING SPEECHLIKE SOUNDS

Astrid van Wieringen and Louis C.W. Pols
Institute of Phonetic Sciences, IFOTT, Amsterdam, The Netherlands

ABSTRACT

Perceptual processing of single and multi-formant CV-like and VC-like sounds, and of natural speech-based syllables is examined in three experiments to determine the extent to which perception of rapid transitions in speech can be explained by general auditory properties. These experiments show that resolution varies with the stimulus complexity and paradigm used, but that it is not controlled by speech-specific properties.

INTRODUCTION

Stop consonants in speech are cued by several properties including short and rapid transitions. Perceptual resolution of such rapid transitions is examined by asking listeners to classify (b or d), discriminate (ABX), and identify (by number 1-7) different kinds of speechlike transitions, which are preceded or followed by a stationary part. It was expected that perceptual processing would depend on the stimulus complexity, and that the number of discriminable or identifiable stimuli would decrease with increasing stimulus complexity: perception would be mediated more by long-term memory and less by acoustical properties with increasing speechlikeness of the stimuli.

STIMULUS GENERATION

Formant synthesis

The single and complex CV-like and VC-like syllables were generated by a digital formant synthesiser [6, 7]. A 110-Hz pulse was used as glottal source. To ensure a precise generation of these formant transitions the stimuli were sampled at 1.2 MHz. After low pass filtering, they were downsampled to 20 kHz (16 bit resolution). The formant frequency values were updated every 1 ms. Although the first period of the stimulus always started on a zero crossing, stimuli were preceded and followed by a 2-ms cosine window to avoid clicks. The formant bandwidth was proportional to the changing formant frequency (10%). The actual stimuli were generated real-time by means of an OROS-AU22 DSP board with D/A converter.

The *single* formant syllables had 30-ms transitions, preceded or followed by 80-ms /a/-like (figure 1) or /u/-like (at 800 Hz) stationary portions. The transitions varied in endpoint frequency from 950 Hz to 1550 Hz in steps of 100 Hz, the average difference limen in frequency for these types of stimuli [5, 6].

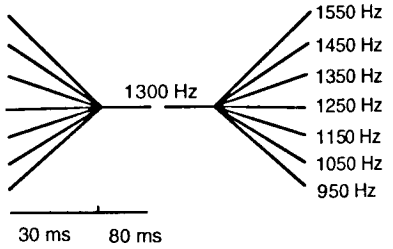


Figure 1. Schematic illustration of initial CV-like (left) and final VC-like (right) /a/-like single formant transitions (not to scale).

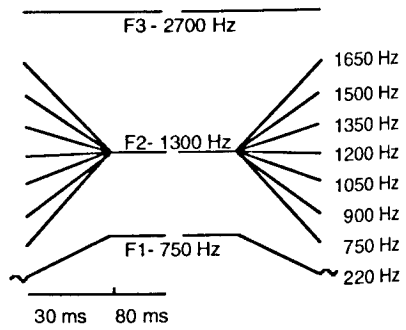


Figure 2. Schematic illustration of initial CV-like (left) and final VC-like (right) /a/-like complex formant transitions (not to scale).

The first-formant and second-formant transitions of the *complex* stimuli were also 30 ms, preceded or followed by an 80-ms /a/-like (1300 Hz) or /u/-like (800 Hz) steady-state. A stationary third formant, and a 20-ms voice bar were added to make the stimuli sound more speechlike (figure 2). The transitions varied in endpoint frequency from 750 Hz to 1650 Hz in 150 Hz steps, the average difference limen in frequency for these types of stimuli [5, 6]. The fixed F1-transitions of the complex syllables rose or fell from 220 Hz to 750 Hz and the F3 was fixed at 2700 Hz for the /a/-like and at 2200 Hz for the /u/-like stimuli.

Interpolated speech-based stimuli

The speech-based stimuli were created by interpolating [4] the spectral envelope of two natural endpoints in seven steps, e.g., /ba/ and /da/. The original /ba/, /da/, /ab/, /ad/ stimuli were segmented from CVC tokens pronounced by a native Dutch male speaker (F0 of about 110 Hz). Stimuli were digitised with a sample frequency of 20 kHz (cut-off frequency of the low-pass filter was 4.9 kHz; slope 96 dB/oct). All syllables were segmented to be 100 ms.

In total twelve seven-syllable continua were created varying along the bilabial-to-alveolar dimension.

PROCEDURE

In the *ABX discrimination* task five subjects were tested individually in a quiet room. Three subjects listened to the /u/-like stimuli, three to the /a/-like ones (one of the subjects listened to both formant patterns). They were seated in front of a terminal and heard three stimuli over Sennheiser headphones at a comfortable level. The inter-stimulus time between the three stimuli was 500 ms. By clicking the appropriate response square on the monitor, they could indicate whether they considered the third stimulus to be identical to the first or to the second, after which three new stimuli were generated. No feedback was given during the test.

After a short training period, each of the four combinations per stimulus pair (ABA, ABB, BAA, BAB) was repeated 25 times, resulting in 100 observations per stimulus pair per subject. All conditions were tested separately. Each test, which was preceded by ten test triads, lasted approximately 10 minutes.

The same listeners also *classified* the single, complex, and interpolated speech-based stimuli as 'b' or 'd' on separate occasions.

In the *absolute identification paradigm* subjects are trained to assign a label (1-7) to each stimulus in a continuum. They have to learn the labels on the basis of their own criteria and therefore use numbers as response labels: no information is given about the nature of the stimuli under test. Feedback is given after each response, to maintain a constant level of performance.

Each subject made a total of 189 responses to the seven stimuli in each stimulus condition. Before each test series there were 63 test trials, which were not taken into account. Fourteen of these test series were collected for each stimulus complexity. Of these the first four were disregarded. Therefore, each of the six stimulus complexities per subject consisted of 1890 responses (270 x 7).

RESULTS

ABX discrimination

Figure 3 illustrates the 1-step ABX-discrimination functions and the classification sigmoids, averaged over the six subjects and two formant patterns (two statistically non-significant factors) together with the predicted discrimination function, based on the average classification sigmoids of these six subjects [2, 6]. The discrimination results are plotted in terms of percentage correct as a function of one pair of stimuli (one pair is averaged over ABA, ABB, BAA, and BAB). The two most striking results are 1) that the predicted and measured functions differ markedly and 2) that categorisation, if any, depends on stimulus complexity and on the position of the transition. As basic sensitivity is comparable within a relatively large frequency range [6], perceptual discontinuities arise from attentional constraints due to the increasing number of cues with increasing complexity and speechlikeness of the stimuli.

In general, subjects discriminate better between subsequent pairs of stimuli than predicted from the classification sigmoids. Compared to the predicted functions, the experimental ones yield

higher percentage correct scores. As for classification performance, figure 3 shows that listeners were indeed able to classify the different kinds of stimuli consistently as 'b' or 'd'.

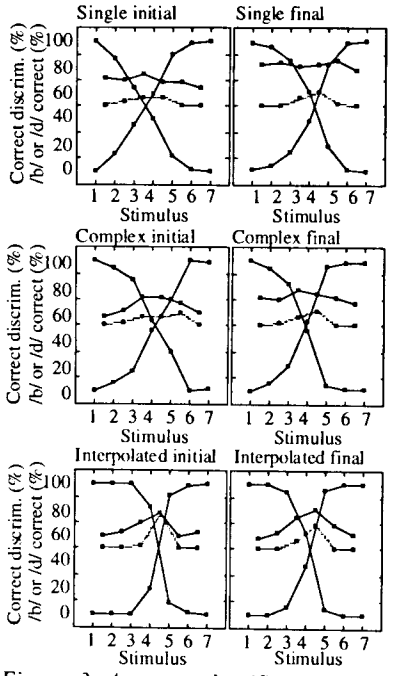


Figure 3. Average classification and discrimination scores (both actual (solid) and predicted (dashed)), averaged over subjects and formant patterns. The stimuli are indicated on the abscissa (the discrimination data apply to pairs of stimuli).

The issue is whether discrimination of single, complex and interpolated stimuli is based on sensory differences or on a phoneme labelling mechanism. From a sensory point of view listeners should be equally sensitive to acoustical differences of the single or the complex formant continua, because the step size is similar in a relative sense (being one JND). However, a different pattern of results is expected if cognitive processes dominate sensory ones: it is then extremely difficult to apply an analytical listening strategy and to differentiate between the different stimuli of the continuum. Our study shows that listeners use acoustical cues to distinguish the seven stimuli of a continuum. Perception of the single formant stimuli can approach the limits of the auditory system, presumably because subjects can listen attentively to the varying acoustical cues. In these conditions the listener is more sensitive to acoustical differences in final than in initial transitions. The more complex the stimuli the more the responses are divided into two categories. However, there is no clear evidence of categorical perception, not even with the interpolated speech-based stimuli. In the case of categorical perception discrimination should be at chance level (50%) for those stimuli that are classified similarly and much higher than chance for those stimuli which are labelled differently.

Absolute identification

To determine whether the complexity of the stimulus induces 'speech categories', the data collected in the absolute identification experiment are also analysed in terms of *d'ident*, the perceptual distance between adjacent stimuli in an absolute identification experiment [1, 6]. Once *d'ident* is computed, the number of categories per stimulus continuum can be determined by means of a criterion. In the case of categorical perception *d'ident* is 1.0 (our arbitrary criterion for indistinguishable pairs of stimuli), for those stimuli which are labelled similarly, and higher than 1.0 for those which are labelled differently.

Figure 4 illustrates performance for the single, complex, and interpolated speech-based stimuli continua in initial (squares) and final (stars) position, averaged over three subjects. Data are plotted in terms of *d'ident* as a function of the neighbouring pairs of stimuli in the continua. The better the stimuli are identified, the smaller the number of confusions, and the higher the *d'ident*. High *d'ident's* were found with the single transitions in final position: subjects are very sensitive to the physical cues of these stimuli, as was also the case in the ABX-discrimination paradigm. The lower the *d'ident*, the less distinguishable the neighbouring pairs of stimuli are.

The figure shows that *d'ident* drops, on average, as the stimuli become more complex, and that the difference between initial and final transitions becomes

smaller with increasing stimulus complexity.

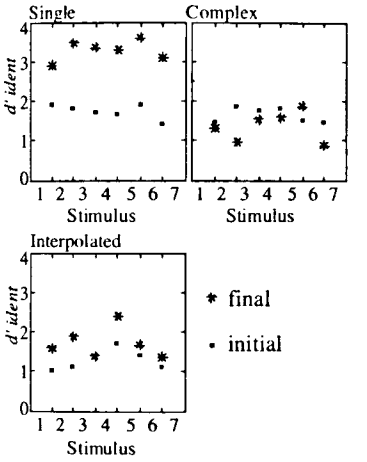


Figure 4. Absolute identification results, averaged over three subjects and two formant patterns. For more details see text.

It was expected that phonemic labelling would restrict the number of categories in a continuum, i.e., that the number of categories would decrease with increasing complexity of the stimulus. This does not appear from our data. As relatively few pairs of stimuli yield a *d'ident* lower than 1.0 we cannot conclude from our data that categorical perception occurs with increasing stimulus complexity. The interpolated speech-based stimuli show increased sensitivity between stimuli four and five, possibly as a result of a phoneme boundary in the stimulus continuum. Our results suggest that listeners perceive the single and complex stimuli in the so-called context-coding mode [3, 6]: they create internal representations of the continua under test and are capable of distinguishing the stimuli within the /b/ and /d/ categories. Although the interpolated speech-based sounds are perceived more categorically than the formant stimuli, the data give no clear evidence that these stimuli are processed by a long-term phoneme-labelling mechanism.

In speech communication listeners (fortunately) do not need to perceive detailed acoustical cues. However, our study shows that the perception of vocalic transitions can, to some extent, be explained by general auditory properties, and that listeners can try to zoom in on certain levels of processing and discriminate ambiguous or new cues if they are not masked. In our study perception does not seem to be limited by a speech-specific mechanism based on long-term linguistic experience. In our study all the stimuli, including the interpolated speech-based ones, are discriminated better than predicted from the 2-AFC classification task, suggesting that listeners make use of additional acoustical cues. Experiments with natural speech transitions showed that the perceptual asymmetry between initial and final transitions decreases with increasing stimulus complexity, presumably because natural speech transitions contain redundant cues for plosive identification [6]. However, further study is necessary to understand how linguistic knowledge influences the perception of vocalic speech transitions.

REFERENCES

[1] Macmillan, N.A. & Creelman, C.D. (1991): *Detection theory: a user's guide* (Cambridge University Press).

[2] Pollack, I. & Pisoni, D.B. (1971). "On the comparison between identification and discrimination tests in speech perception", *Psychonomic Science* 24, 299-300.

[3] Schouten, M.E.H. & Van Hessen, A.J. (1992): "Modeling phoneme perception. I: Categorical perception", *Journal of the Acoustical Society of America* 92, 1841-1855.

[4] Van Hessen, A.J. (1992): "Discrimination of familiar and unfamiliar speech sounds", *Ph.D.-thesis*, Univ. of Utrecht.

[5] Van Wieringen, A. & Pols, L.C.W. (accepted): "Discrimination of single and complex CV- and VC-like formant transitions", *Journal of the Acoustical Society of America*

[6] Van Wieringen, A. (1995): "Perceiving dynamic speechlike sounds: psychoacoustics and speech perception", *Ph.D.-thesis*, University of Amsterdam.

[7] Weenink, D.J.M. (1988): "Klinkers: een computerprogramma voor het genereren van klinkerachtige stimuli", *IFA-report nr. 100.*

ACOUSTIC SEGMENTATION USING VARIOUS NEURAL ADAPTATION MODELS

E. Jones, University College Galway, Ireland.
E. Ambikairajah, Regional Technical College, Athlone, Ireland.

ABSTRACT

This paper describes initial results of a comparison of several published models for the inner hair cell/auditory-nerve synapse, for the task of speech segmentation. In each case, the hair cell/synapse model is combined with a model for basilar membrane filtering, and a segmentation algorithm is applied to the neural firing rate in order to emphasize the acoustic boundaries in the speech. The models are tested using utterances from the TIMIT acoustic-phonetic corpus. Performance of each model is assessed by comparing the segmentation it produces with the phonetic transcription provided with the TIMIT database.

1. INTRODUCTION

Segmentation is an important step in the mapping of an acoustic speech signal to a lexicon of word or sub-word units. This is useful in a number of applications, including computer recognition of continuous speech and transcription of speech corpora.

Several researchers have combined the operations of phonetic segmentation and classification into one step. For example, [1] describes a hidden Markov model-based system which classifies a set of 48 phones from the TIMIT database. This system uses a priori knowledge of the phonetic or orthographic content of the utterance to provide phone boundaries. A similar system for phone classification is described in [2].

Another technique which has been used for segmentation is the extraction of a 'boundary function' from a spectral or parametric representation of the utterance. This function encapsulates information about the acoustic boundaries, and can be used either to assist in classification-based segmentation by providing additional temporal information, or as a starting point for a separate stage of classification. For example, in [1], a 'Spectral Variation Function' was extracted from a mel-cepstral representation of the speech to provide additional information about acoustic transitions. Alternatively, in [3], an 'association strength' provided initial potential boundaries, which were subsequently used to generate a multi-level segmentation (a 'dendrogram'), from which final phonetic boundaries were derived using a search procedure.

The boundary function used in [3] was derived from a spectral representation provided by an auditory model [4]. This was found to give better performance, in terms of the least number of boundary insertion and deletion errors, than other representations including discrete Fourier transform and linear predictive coding. One of the reasons for this better performance is the fact that the auditory model used ([5]) included a model for the inner hair cell/auditory-nerve synapse, which exhibits adaptation and recovery, i.e. it enhances sudden onsets and offsets.

While the work described in [4] compared an auditory representation to non-auditory based representations, little work has been carried out to assess the segmentation performance of some other published models for neural adaptation. This issue is addressed in this paper where several adaptation models are combined with a computational model for basilar membrane (BM) mechanics, and used to process continuous speech utterances from the TIMIT database. The representation produced by each model is further processed to produce a time function which contains markers corresponding to potential acoustic boundaries. Performance evaluation is carried out by comparing the auditory-derived boundary markers with those provided with the TIMIT database.

2. THE AUDITORY MODELS

2.1 BM Model

The model for BM mechanics used in this study is based on the transmission line model described in [6] and [7], and consists of a cascade of 128 digital filters covering the frequency range from 70 Hz to 3.4 kHz. The sampling frequency is 8 kHz. The output of each filter in the cascade, corresponding to BM displacement, is used as input to each IHC/synapse model.

2.2 Adaptation Models

The adaptation models chosen for this study were:

- the IHC/synapse model of Seneff [5];
- the Schroeder-Hall reservoir model, as implemented by Cohen [9];
- the adaptation model of Meddis [8];
- an alternative model, based on Meddis's model (Jones at al. [7]).

Since an auditory model could be a useful component of a practical continuous speech recognition system, special emphasis was placed on the adaptation models' relative computational complexity. To this end, it was decided to modify the adaptation models to operate at the same sampling frequency as the BM model, 8 kHz. Both Cohen's model and the alternative model of [7] have particularly simple structures which could be a useful advantage from the point of view of computational load.

As the models operated at 8 kHz, the speech utterances from the TIMIT database, which have a sampling frequency of 16 kHz were decimated by 2. The downsampled speech was processed by each composite BM/adaptation model, with a single neural firing rate vector produced every 5 ms.

3. SEGMENTATION ALGORITHM

The segmentation algorithm applied to the sequence of neural firing rate vectors is based on that described in [4], and operates on the assumption that speech segments can be distinguished from each other by measuring the differences between their spectral representations. The algorithm starts at the beginning of the sequence of spectral vectors and 'associates' each frame with either its past or its future, based on the cumulative distance between that frame and its immediate backward and forward neighbours, over a certain observation range. This observation range was set equal to 50 ms on either side of the current frame [4]. Each frame, n, accumulates forward and backward distances $D_f(n,k)$ and $D_b(n,-k)$, between itself and neighbouring frames k, where $D_f(n,k)$ is defined as follows:

$$D_f(n,k) = \sum_{j=0}^{k} d(n,j)$$

where $d(n,j)$ denotes the Euclidean distance between frame n and frame $n+j$. $D_b(n,-k)$ is defined in a similar manner.

Forward and backward distance measures are accumulated in parallel until either the difference between them exceeds a certain minimum distance, D_{min}, or the observation range is exceeded. An 'association strength', which is the maximum difference between D_f and D_b over the association range, is assigned to each frame. The association strength contour is smoothed using a Gaussian window with a variance of 5 ms [4].

An example of the association strength contour for a portion of the TIMIT utterance "she had your dark suit in greasy wash water all year" is shown in Fig. 1(a). This figure gives the contour derived from Seneff's adaptation model (for clarity, the contours are displayed only as far as the word "suit"). The positive-to-negative crossings of the contours indicate the location of potential acoustic boundaries; this information is converted into a series of pulses, with the height and width of the pulses corresponding to the strength and abruptness of the acoustic change (Fig. 1(b)). A threshold is applied to the pulse sequence, such that pulses below this threshold are set equal to zero. Thus, small pulses which may correspond to false acoustic boundaries are eliminated [4].

The final stage of the segmentation process is the conversion of the pulse sequences of Fig. 1(b) into binary sequences where a '1' indicates the presence of an acoustic boundary at a particular frame, and a '0' indicates no boundary. This allows a straightforward comparison between the boundaries produced by the adaptation models and those obtained from the TIMIT transcriptions. Figure 2 displays such binary sequences for all four adaptation models, where the boundaries provided by the TIMIT transcription are indicated by the pulse train in part (e) of the figure.

4. PERFORMANCE EVALUATION

Initial segmentation parameter choice and performance evaluation was carried out using a subset of the TIMIT database consisting of 40 sentences, with over 1400 boundaries. The binary pulse train produced by each BM/adaptation model combination was compared with the pulse train extracted from the TIMIT transcription and the number of alignments, deletions and insertions was noted. Clearly, the performance of the system as a whole depends as much on the choice of parameters for the segmentation algorithm as it does on the representation produced by the auditory front-end. Since some parameters affect the number of insertions and deletions in different ways, an 'equal error' criterion was used for choosing certain parameters, i.e. the value which gave equal numbers of insertions and deletions was chosen. The numbers of alignments and errors were then used as a measure of relative performance.

Table 1 summarises the relative performance of the models examined, where a tolerance of 25 ms was used for boundary alignment. While the equal error criterion is useful in the current study, the absolute performance of any given model could be improved upon, by considering additional strategies for reducing the number of insertions or deletions, e.g. in a further classification stage it might be possible to recover some deleted boundaries [4], therefore at the segmentation stage a smaller number of insertions can be achieved at the expense of a larger number of deletions.

Table 1. Summary of segmentation performance of all four adaptation models (total errors = sum of insertions and deletions).

Model	Total alignments	Total errors	% Correct
Seneff	1045	842	72.4
Cohen	1112	667	77.1
Meddis	970	962	67.2
Jones et al.	1086	699	75.3

5. DISCUSSION AND CONCLUSIONS

This paper has presented initial results from a comparison between various neural adaptation models, applied to the task of determining acoustic boundaries in continuous speech. The segmentation method used does not make use of any a priori knowledge of the phonetic sequence, it relies solely on the information extracted from the speech by the auditory front-end processor. From experiments with a subset of sentences from the TIMIT database, it would appear that Cohen's model provides slightly better performance than the model of [7], with the other two models a few percent behind. It is interesting to note that the two models which present the lightest computational load give the best performance. Future work will involve validation of these results using a larger database, as well more detailed analysis of the nature of the segmentation errors which occur.

REFERENCES

[1] Brugnara, F., Falavigna, D. & Omologo, M. (1993), "Automatic segmentation and labeling of speech based on hidden Markov models", *Speech Comm.* 12, pp. 357-370.
[2] Ljolje, A. & Riley, M. (1991), "Automatic segmentation and labeling of speech", *Proc. ICASSP*, pp. 473-476.
[3] Glass, J. & Zue, V. (1988), "Multi-level acoustic segmentation of continuous speech", *Proc. ICASSP*, pp. 429-432.
[4] Glass, J. & Zue, V. (1986), "Signal representation for acoustic segmentation", *Proc. SST-86*, pp. 124-129.
[5] Seneff, S. (1988), "A joint synchrony/mean-rate model of auditory speech processing", *J. of Phonetics* 16, pp. 55-76.
[6] Ambikairajah, E. Black, N. & Linggard, R. (1989), "Digital filter simulation of the basilar membrane", *Comp. Speech & Lang.* 3, pp. 105-118.
[7] Jones, E. & Ambikairajah, E. (1993), "Comparison of various adaptation mechanisms in an auditory model for the purpose of speech processing", *Proc. Eurospeech '93*, pp. 717-720.
[8] Meddis, R. (1986), "Simulation of mechanical to neural transduction in the auditory receptor", *J. Acoust. Soc. Am.* 79, pp. 702-711.
[9] Cohen, J. R. (1989), "Application of an auditory model to speech recognition", *J. Acoust. Soc. Am.* 85, pp. 2623-2629.

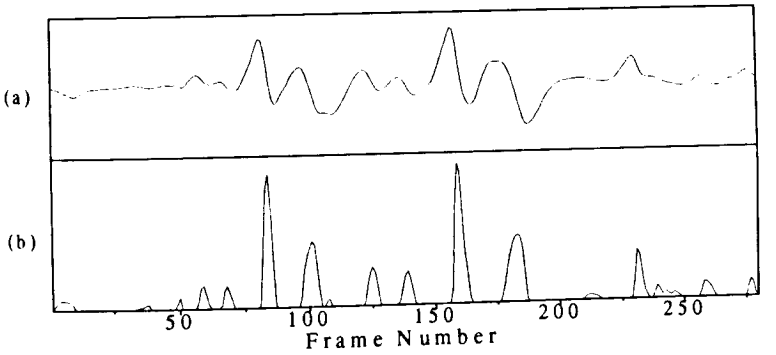


Figure 1. (a) Association strength contour, and (b) boundary pulse sequence for the TIMIT utterance fragment "She had you dark suit" obtained using Seneff's adaptation model.

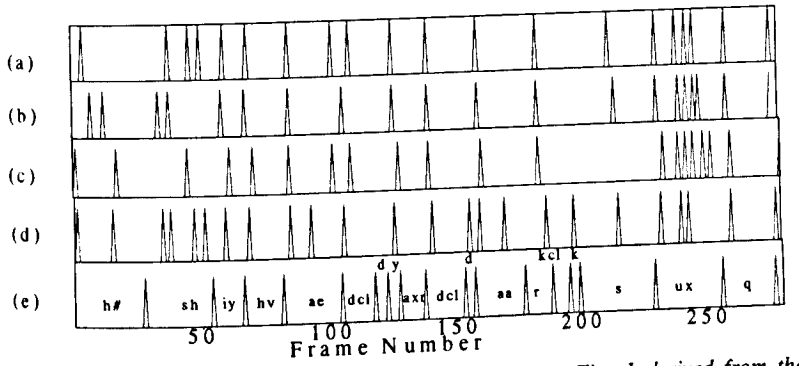


Figure 2. Boundary markers for the same utterance as in Fig. 1 derived from the following adaptation models: (a) Seneff (b) Cohen (c) Meddis, and (d) Jones et al. [7]. The actual boundaries derived from the TIMIT transcription are given in part (e), with phonetic labels.

DISCOURSE-BASED EMPIRICAL EVIDENCE FOR A MULTI-CLASS ACCENT SYSTEM IN FRENCH.

C. Astesano, A. Di Cristo & D.J.Hirst
Institut de Phonétique, CNRS URA 261 "Parole et Langage"
Université de Provence, France

ABSTRACT

In this paper we present a pilot study of French accentuation in discourse, based on two types of corpora, consisting of FM broadcast news and interviews. The main purpose of this research is to validate empirically, by means of a pluriparametric analysis of raw and normalised data, the accentual categories which are considered as constitutive of the contemporary French accent system.

INTRODUCTION

There are two current views about French accentuation : the *extreme view*, as a language without accent ([1], for a discussion) and the *traditional view*, which claims that French possesses only two types of accents : a phrase-final rhythmic accent and a word-initial emphatic accent [2, 3]. This view is extremely frequent in recent works on prosodic phonology [4, 5, 6]. Data gathered in recent studies, however, which have been carried out on different corpora involving various speech styles [7, 8], together with results of our own work [9], strongly suggest that these two points of view are questionable and lead us to conclude that accentuation in French is far more complex than has previously been stated. To explain the peculiar nature of this accentuation, we propose a general typological framework based on three main classes of accents : lexical, rhythmic and semantico-pragmatic. It is well known that French is only concerned by the last two classes.With regards to the rhythmic class, besides the traditional *phrase-final accent*, modern French is characterised by an optional *secondary accent* which is assigned to the first syllable of a polysyllabic word in order to avoid a accent clash (i.e."un *cha*peau blanc") or more generally to favour the formation of an eurhythmic pattern. Despite its early identification [7], this secondary accent has been neglected by phoneticians and only recently has been incorporated into a phonological description of French prosody [10].

The semantico-pragmatic accent class can be divided into two sub-categories: the non-emphatic class, which contains the *nuclear accent* (associated with the tonic of an Intonation Unit) and the emphatic class which includes both the contrastive accent and the focal accent for intensification usually named *accent d'insistance* . The latter, which is most often assigned to the first syllable of a word, is often confused in the litterature with the secondary rhythmic accent which occupies the same position. It has been argued, nevertheless [11], that the secondary accent is a true pitch accent, meaning that its realisation is not accompanied by any lengthening effect, a feature which has been assumed to be common both to the primary phrasal accent, the nuclear accent and the two emphatic accents. It is the purpose of this paper to verify if this assumption is tenable for discourse and to what extent we can associate each of the categories of accent proposed above with specific qualitative parameter features.

A second aim of this study is to examine the way in which lengthening is distributed throughout the syllable for the different accent classes we define. Campbell [12] suggested the strong hypothesis that once raw duration values of phonemes have been normalised as z-scores, the presence of a lengthening effect is distributed equally throughout the syllable in English. Bailly & Barbosa [13] on the other hand claimed that the relevant unit for lengthening in French is not the syllable but the sequence of phonemes from one vowel onset to the next. Other studies, however, have suggested that lengthening does not apply equally to the different syllable constituents. Thus Fant & Kruckenberg [14] for example found that postvocalic consonants in French were lengthened only in prepausal position. Similar

results were reported for English by Campbell [15] who found that codas were lengthened more in pre-boundary position while onsets were lengthened more in prominent syllables not followed by a boundary.

MATERIAL AND PROCEDURE

We extracted from our database on prosody two FM recordings of continuous speech lasting approximately 2 minutes each by two native male speakers of educated standard French. The first was an extract from a radio news broadcast and the second an interview. These recordings were transcribed without punctuation. As a preliminary test three experts were asked to indicate all perceived accents, to mark emphatic accents and non-terminal and terminal Intonation Unit boundaries. Approximately 330 accents were identified. These were classified into the following categories : emphatic (EMP), final in a terminal Intonation Unit (IU-T), final in a non-terminal Intonation Unit (IU-N), word-initial (WI), and (prosodic) word-final (WF). The remaining syllables were labelled as unaccented (UN).

Experimental procedure:

The excerpts were digitalized at 16 kHz on a Sun Sparc station and labelled phonemically and in syllables by hand. Approximately 2500 phonemes and 1200 syllables were labelled.

Each constituent of the syllable was coded as onset, nucleus or coda.

Duration. The duration of the different syllable constituents was measured and the raw data was normalised using the Z transform method [12], with the phonemic means and standard deviations pooled from a database of seven speakers.

Fundamental frequency. The fundamental frequency of the extracts was modelled with a quadratic spline function using an automatic modelling algorithm Momel with manual corrections [16]. For each syllable the nearest maximum target was calculated as well as the distance of the target with respect to the onset of the corresponding vowel. F0 values were normalised using the ERB scale [17] offset to the mean of the speaker's range.

RESULTS

Duration

Analysis of variance on the phoneme durations showed considerable differences for the different accent classes. Both accent class and position in the syllable were highly significant factors ($p < 0.0001$) and the interaction between the two factors was also highly significant ($p < 0.0001$). These effects were observed for both speakers on both raw and normalised durations. Figure 1 shows the normalised durations for Onset, Nucleus and Coda for the different accent-classes for speaker 1. It can be seen that the ratio between nucleus and coda remains fairly constant showing that the syllable rhyme seems to be treated as a consistent unit for lengthening in different environments.

The syllable Onset, by contrast, behaves rather differently. The onset is longer than the other constituents of the syllable in unaccented syllables and word-initial accents, both emphatic and non-emphatic. The onset is shorter than the other syllable constituents when the syllable occurs at the end of either a Terminal or a Non-Terminal Intonation Unit. When the syllable was word final but not at the end of an Intonation Unit the ratio of the onset to the other syllables was intermediate.

Results for the second speaker showed similar effects except that the relationship between the Nucleus and the Coda were less constant. The ratio of the syllable Onset to the rhyme showed the same effects as for the first speaker in the different contexts (Figure 2).

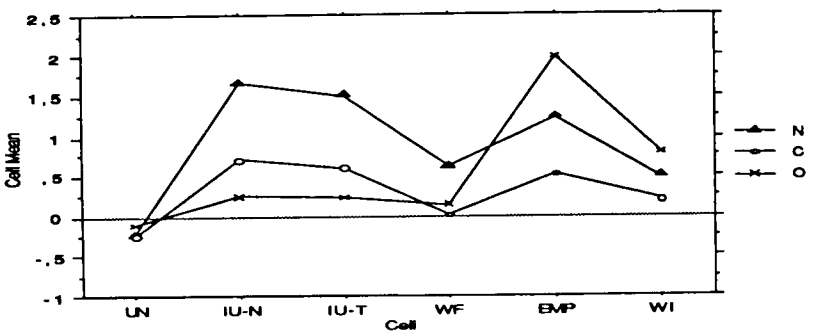


Figure 1 : Mean durations of onset (O), nucleus (N) and coda (C) for the different accent classes (see text) for speaker 1.

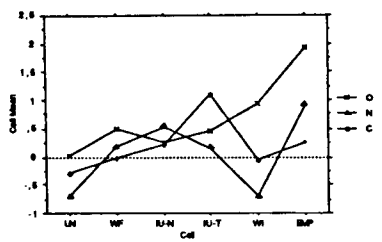


Figure 2 : Mean durations of onset (O), nucleus (N) and coda (C) for the different accent classes (see text) for speaker 2

Fundamental frequency

Figure 3 shows the mean target values in ERB by accent class for speaker 1. Analysis of variance showed that the differences between the classes were highly significant ($p < 0.0001$). Post hoc tests confirmed that each pair of values was significantly different ($p < 0.05$) except between the unaccented syllables and the syllables in final position in Terminal Intonation Units.

The temporal location of the target points was also significantly different for certain of the classes, being located at a mean of 91 ms from the vowel onset for the end of Non-Terminal Intonation Units, 66 ms for emphatic accent and from 30-40 ms for the other accent classes which were not significantly different from one another.. Similar results were observed for the second speaker.

DISCUSSION

The different categories of accent type which we hypothesised were all clearly distinguished by the acoustic parameters. Duration was particularly effective in distinguishing preboundary accents from others confirming for both read and spontaneous speech results mentionned above [14, 15]. The only category where durational effects did not play a role was the distinction between Terminal and Non-Terminal Intonation Units which were distinguished by the value of the F0 target and its timing (cf Vincent et al. [18] for similar findings).

These results can be summarised in the following table where the value of duration is represented as very short (--), short (-), long (+) or very long (++).

	Onset	Rime	F0	Timing
UN	--	--	low	early
WI	+	+	high	early
EMP	++	++	high++	mid
WF	-	+	high	early
IU-N	-	++	high+	late
IU-T	-	++	low	early

The results of this preliminary study suggest that the parameters we have identified characterizing the different accent classes are relatively independant of discourse type.

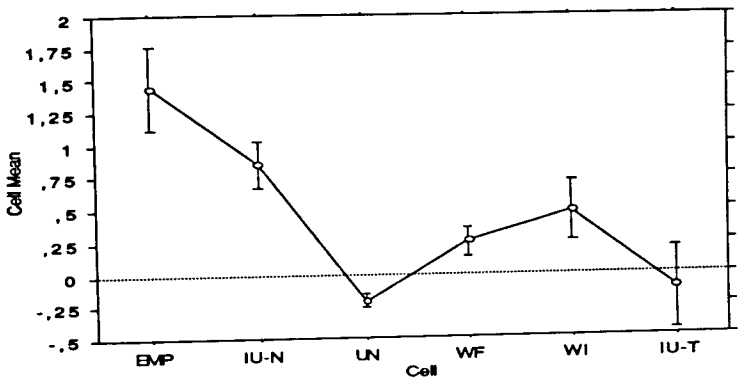


Figure 3 : Target points (in ERB units) for the different accent classes (see text) for speaker 1 . Vertical line indicate 95% confidence intervals.

REFERENCES

[1] Rossi, M., (1980), "Le français, langue sans accent? Problèmes de Prosodie.", *Studia Phonetica*, vol.15, pp. 13-51.

[2] Delattre, P. (1938), "L'accent final en français: accent d'intensité, accent de durée", *French Review* , vol.12(2), pp. 141-145.

[3] Séguinot, A., (1976), "L'accent d'insistance en français standard", *Studia Phonetica* , pp.1-91.

[4] Hyman, L., (1975), Phonology: Theory and Analysis., Holt, Rinehart & Winston, N.Y.

[5] Halle, M. & Vergnaud, J.-R., (1987), *An essay on stress*, Cambridge, Mass.

[6] Fletcher, J. (1991), "Rhythm and final lengthening in French", *Journal of Phonetics*, vol.19, pp.193-212.

[7] Fónagy, I (1980), "l'accent en français: accent probabilitaire", *Studia Phonetica* , vol.15, pp.123-233.

[8] Vihanta, V., (1993), "Focalisations et autres proéminences en français lu et spontané", *Mélanges Lingren*, Turun Yllopisto, pp.258-289.

[9] Hirst, D.J. & Di Cristo, A. (1984), "French intonation: a parametric approach". *Die Neueren Sprachen*, vol.83, pp.554-569.

[10] Di Cristo, A. & Hirst, D.J. (forthcoming), "l'accent en français: stratégies et paramètres", (*Hommages Fónagy*)

[11] Rossi, M., (1985), "L'intonation et l'organisation de l'énoncé", *Phonetica*, vol.42, pp. 135-153.

[12] Campbell, W.N., (1992), *Multi-level Timing in Speech*, PhD Thesis, University of Sussex.

[13] Barbosa, P. & Bailly, G., (1994), "Characterization of rhythmic patterns for text-to-speech synthesis", Speech Communication, vol.15, pp.127-137.

[14] Fant, G.; Kruckenberg, A.; Nord, L., (1991), "Durational correlates of stress in Swedish, French and English", *Journal of Phonetics*, vol. 19, pp. 351-365.

[15] Campbell, W.N., (1993), "Automatic detection of prosodic boundaries in speech", *Speech Communication*, vol.13, pp. 343-354.

[16] Hirst, D.J. & Espesser, R. (1993), "Automatic modelling of fundamental frequency with a quadratic spline function." *Travaux de l'Institut de Phonétique d'Aix*, vol. 15, pp. 71-85.

[17] Hermes, D.J. & Van Gestel, J.C. (1991), "The frequency scale of speech intonation.", *J. Ac. Soc. of Am.*, vol. 90, pp. 97-102.

[18] Vincent, M.; Di Cristo, A. & Hirst, D.J., (1995), "Prosodic features of finality for intonation units in French discourse" (These proceedings).

ARTICULATORY CORRELATES OF SECONDARY STRESS IN POLISH AND SPANISH

Gabriele Scharf[1], Ingo Hertrich[1], Iggy Roca[2], Grzegorz Dogil[3]
1 Department of Neurology, University of Tübingen, Germany
2 Department of Language and Linguistics, University of Essex, England
3 Institute of Computational Linguistics, University of Stuttgart, Germany

ABSTRACT

Electromagnetic Articulography was used to registrate tongue movements during Spanish and Polish sentence utterances. In Polish no significant differences between secondarily stressed and unstressed syllables could be found. In Spanish the tongue gesture of the secondarily stressed syllable showed a different shape of the movement in comparison with the unstressed case. However, this effect might represent a coarticulatory artifact of the test material.

INTRODUCTION

An alternating secondary stress for Polish and Spanish has been postulated by numerous phonologists although the phonological as well as the acoustic phonetic evidence for an alternating secondary stress in Polish and Spanish is tenous, if not non-existent. There are no phonological processes or rules that would crucially refer to the position of secondary stress. There can't be found clear acoustic correlates of an alternating secondary stress such as F0-, intensity, durational or spectral features differing from the unstressed syllables in both languages either (cf. [6], [3]).

DeJong et al [2] found that in English sentence utterances stress on different prosodic levels was implemented by the same articulatory feature, namely movement size as compared to the unstressed case, but to a different extent: The movement amplitudes of the secondarily stressed syllable showed values just in between the amplitudes of the nuclear stressed and the unstressed syllable. However, Beckman & Edwards [1] suggest that, at least in English, stress on different prosodic levels yields different phonological and phonetic - and articulatory - realisation: They found that nuclear stress was mainly correlated with intonational features whereas at lower prosodic levels like word level stress was correlated with longer syllable duration as well as with articulatory features like greater movement amplitude and greater velocity.

The articulatory realisation of secondary stress in Polish and Spanish has not been investigated before. The purpose of the present study was to find an articulatory correlate of alternating stress in Polish and Spanish sentence utterances and to answer the question wether secondary stress is implemented by the same articulatory feature as primary accent - but to different degrees - or by a principally different kind of feature.

METHODS

In the present experiment the Articulograph AG100 (Carstens Medizinelektronik Göttingen, cf. [9]) was used. The kinematic recordings were made with a sampling rate of 200 Hz. Simultaneously with the recording of the articulatory data the acoustic signal was digitally recorded. Five sensors were placed on upper lip (UL), lower lip (LL), 5mm behind the tongue tip (TT), tongue mid (TM) and tongue dorsum (TD). Different word forms of Polish and Spanish stimulus words were used with one target syllable appearing on three different prosodic levels (primary stress, secondary stress, unstressed). The target words for Polish were: *,hipo'potam* (hipopotamus, nom. sg.), *hipopo'tama* (gen. sg.) and *,hipo,pota'mami* (instr. pl.) with the second po being the target syllable. The words were embedded in a neutral sentence frame: *On powie _ dwa rasy* (He'll say _ twice). The Spanish target words were: *Constan'tino* (Constantine), *constan,tino'pleno* (Constantinople man) and the infinitive form of the verb *cons,tanti,nople'ar* (to hang out in Constantinople) with the syllable *ti* being the target syllable. The Spanish carrier sentence was: *No he dicho _ jamas* (I never said _). Each sentence type was visually presented 8 times in randomized order. The speakers, one native speaker of peninsulan Spanish and one native speaker of Polish (two of the authors, I.R. and G.D.) read the sentences with normal speaking rate.

During the production of the sequence *hipopo* in the Polish sentence the mid tongue performs a continuant back- and downward movement from the high position for the vowel [*i*] over the first *po* to the vocal tract configuration of the 2nd [*o*], during the target syllable turning back for producing the following closure [*t*] (as schematized in Fig.1a).

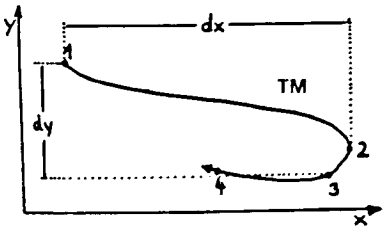


Figure 1a. Schematized mid tongue trajectory during the production of the sequence ipopo in the Polish utterance.

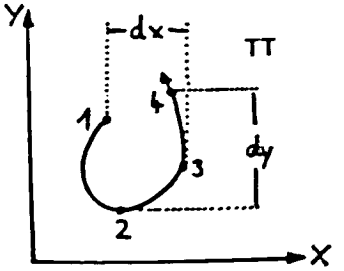


Figure 1b. Schematized tongue tip trajectory during the production of the target syllable ti in the Spanish utterance.

The articulation of the Spanish target syllable is mainly realised by the tongue tip. In the two-dimensional, midsagittal representation the tongue tip performs a circle-like movement from the alveolar closure during the [*t*] over the opening of the vocal tract for the vowel [*i*] to the next alveolar closure for the [*n*] in the following syllable (as indicated in Fig.2a). Across all recorded Polish and Spanish utterances the characteristic articulatory positions (maximum high/front, back and low tongue position) show a certain time alignment with acoustic events (mid of stop consonant closure, onset, mid and offset of the vowel, see the 4 marks in the acoustic signal, cf. Figures 2a, 2b corresponding to the 4 marks in the articulatory trajectory, cf. Figures 2a, 2b). At these characteristic 4 timepoints the x- and y-coordinates of the relevant sensor were calculated. The differences between the extrema on the horizontal and vertical axis (dx, dy) were the basic movement parameters. In addition the following articulatory parameters were derived: the sum dxy = dx + dy as a rough measure for the size of the movement and the quotient dq = dy / dx as a measure for the relation between horizontal and vertical movement and therefore as a rough measure characterizing the shape of the movement. The following durational values were measured in the acoustic signal: for the Polish sentences the vowel duration of [*o*] in the target syllable was extracted as well as the duration of the whole sequence *ipopo*, for the Spanish material the syllable duration of the target syllable *ti* (mid of consonant closure in [*t*] to vowel offset of [*i*]) was measured.

RESULTS

Durations - Polish and Spanish

For both languages the durational values showed a significant difference between main stressed and unstressed syllables but no significant difference between the unstressed syllable and the one with supposed secondary stress (*ANOVA*, alpha=.05). That is, in the present sentence utterances primary stress was realized by longer (vowel/syllable) durations in comparison to the unstressed case whereas secondary stress did not.

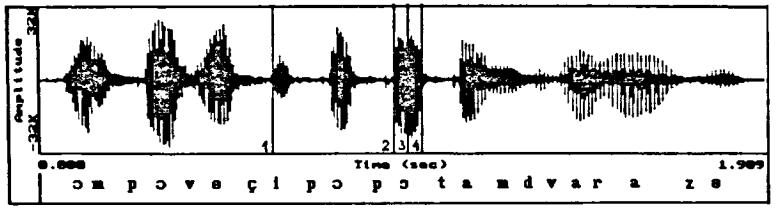


Figure 2a. Acoustic signal of one Polish test sentence with 4 measurement points marked: 1: Vowel onset [i], 2: Vowel onset of 2nd [o], 3: Mid of vowel of 2nd [o], 4: Vowel offset of 2nd [o].

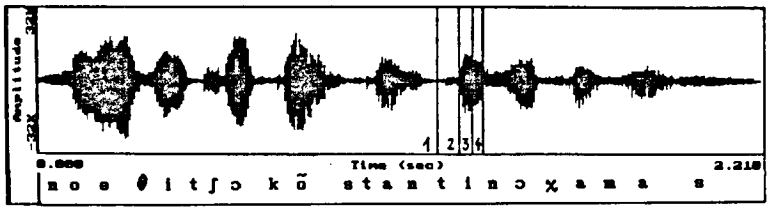


Figure 2b. Acoustic signal of one Spanish test sentence with 4 measurement points on the target syllable marked: 1: Mid of stop consonant closure of [t], 2: Vowel onset [i], 3: Mid of vowel [i], 4: Vowel offset of [i].

Articulatory data - Polish:

In the case of Polish the primary stressed syllables showed significant larger movement amplitudes as compared to the unstressed case but the secondary stress condition did not differ significantly from the no stress case, neither with respect to movement size nor to movement shape (alpha=0.05 in *MANOVA*).

Articulatory data - Spanish:

Figure 3 shows the horizontal and vertical tongue tip movement amplitudes during the production of the syllable *ti*: In both dimensions larger amplitudes were registrated on the target syllable with primary stress than on the secondary stressed or unstressed syllable. The secondary stressed and the unstressed syllable did not differ in movement size. However, they differed slightly with respect to the relation between the two dimensions of the movement: the unstressed syllables showed a small tendency to larger values on the vertical dimension and the syllables with (supposed) secondary stress showed slightly larger horizontal movements. Principal component analysis revealed movement *size* and movement *shape* as the two relevant factors which differentiate between the three accent categories. Post hoc tests showed that the size effect of the primary accent as well as the shape

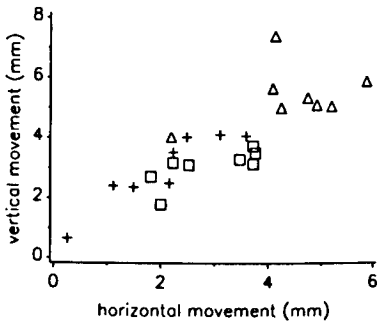


Figure 3. Maximum amplitudes of the tongue tip movement on the horizontal and vertical dimension during the production of the target syllable ti (each symbol represents the measured value of the target movement of one sentence). Primary stress: triangles, secondary stress: squares, no stress: crosses.

effect of the secondary accent were significant ($p<.05$) whereas the difference in movement size between secondary stressed and unstressed syllables did not reach significance ($p>.1$).

DISCUSSION

In the present study longer (syllable) durations and larger articulatory movements of syllables with primary stress as compared to unstressed syllables were measured for both Polish and Spanish sentence utterances. This confirms previous results on articulatory correlates of stress in English (cf. [1], [2], [5]). The assumption that alternating secondary stress would be implemented in the same way as primary stress but to a lower degree could not be confirmed for the material of the present study. For the Polish utterances no articulatory correlate of secondary stress could be found. This negative result might be due to the extremely limited data of the present study with respect to number of speakers and utterances as well as articulatory parameters so that it can't be claimed that secondary stress in Polish does not exist at all.

In the case of Spanish an effect of movement shape on the secondary stressed syllable could be observed. However, since the segmental content of the three target words was not absolutely identical (*Constantino, constantino-pleno, constantinoplear*) it cannot be excluded that the observed small shape effect was caused by the different segmental context rather than by the secondary accent. Nevertheless the strong hypothesis would be the following: Whereas primary stress is implemented by larger articulatory movements secondary stress in Spanish affects the movement shape. This result would push the idea formulated by [1] that stress on different levels might yield different phonetic - and articulatory - implementation. To verify this hypothesis it is necessary to use reiterant speech in future experiments to exclude effects of different contexts and different word length.

Whereas the effect of primary stress on movement size has been explained by greater sonority which may be achieved by e.g. larger downward movements of the jaw (cf. [4]) an explanation for a shape effect of secondary stress is not obvious. One - admittedly very vague - idea might be that a rhythmic variation of movement shape is used as a sort of economization strategy, possibly interpretable as an application of the Obligatory Contour Principle *OCP* of nonlinear phonology (cf. [7]), which does not allow two adjacent identical elements, on the level of articulation.

REFERENCES

[1] Beckman, M.E. & J. Edwards (1994), Articulatory evidence for differentiating stress categories. In: Keating, P.A. (ed.): *Phonological structure and phonetic form. Papers in Laboratory Phonology, vol. III*, Cambridge: University Press.

[2] De Jong, K., M.E. Beckman & J. Edwards (1993), The interplay between prosodic structure and coarticulation. *Language and Speech*, vol. 26, pp. 197-212.

[3] Dogil, G. (in press), The phonetic manifestation of word stress. In: Van der Hulst, H. (ed.): *Word prosodic systems of European languages*. Berlin: De Gruyter.

[4] Edwards, J. & M.E. Beckman (1988), Articulatory timing and the prosodic interpretation of syllable duration, *Phonetica*, vol. 45, pp. 156-174.

[5] Edwards, J., M.E. Beckman & J. Fletscher (1991), The articulatory kinematics of final lengthening, *Journal of the Acoustical Society of America*, vol. 89, pp. 369-382.

[6] Prieto & Van Santen (1995), Acoustic cues of secondary stress in Spanish, Manuscript, Murray Hill: AT&T Bell Laboratories.

[7] McCarthy, J. (1988), Feature geometry and dependency: a review, *Phonetica*, vol. 43, pp.84-108.

[8] Scharf, G., I. Hertrich, G. Dogil & I. Roca (in press), Articulatory correlates of secondary stress in Polish and Spanish, *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS)*, University of Stuttgart.

[9] Schönle, P.W. (1988), *Elektromagnetische Artikulographie. Ein neues Verfahren zur klinischen Untersuchung der Sprechmotorik*, Heidelberg: Springer.

PHONETIC EVIDENCE FOR PHONOLOGICAL STRUCTURE: WORD STRESS IN LATVIAN

A. Krišjānis Kariņš
University of Pennsylvania

ABSTRACT

Goldsmith [1] and Halle & Vergnaud [2] assume that Latvian has only word-initial stress, and cite it as an example of a language with left-headed unbounded feet. Traditional grammars such as Endzelins [3] claim that Latvian words have secondary as well as primary stress. This paper provides phonetic and phonological evidence substantiating the claims of traditional grammars that Latvian has secondary stress.

SEGMENTAL DURATION AND STRESS

Latvian has a system of phonemically contrasting long and short vowels [4]. Bond [5] shows that the duration ratio between phonemically long and phonemically short vowels is approximately 2:1. She also shows that stressed vowels are longer in duration than their unstressed counterparts. Among consonants, phonemic length contrasts exist only for sonorants. Phonemically lengthened sonorants or "lexical geminates" occur in relatively recent loanwords, such as *panna* 'pan' and *ķemme* 'comb'. There is no phonemic length distinction for the obstruents.

In almost all Latvian words, primary stress occurs on the first syllable. The exceptions are words with primary stress on the second syllable [3, 4]. There is no dispute about primary stress in the linguistic literature. The disagreement lies in the differing claims concerning secondary stress.

EXPERIMENT

Following the suggestion of Hayes [6], I consider external phonological evidence for secondary stress. Laua [4] writes that in Latvian, voiceless obstruents lengthen phonetically following stress between two phonemically short vowels. Based upon this assertion, I designed an experiment to answer the following questions: (1) Does quantitative descriptive phonetics provide evidence that consonants are lengthening where they are predicted to lengthen? (2) Is lengthening restricted to the voiceless consonants? (3) Is lengthening restricted to the environment between a short stressed vowel and a short unstressed vowel? (4) Does the distribution of lengthened consonants provide any evidence for secondary stress?

The experimental stimuli were 39 words containing various consonants in positions where they would and would not be expected to lengthen. The words were placed in a neutral carrier phrase, repeated 10 times, and randomized. The resulting list of sentences was read by two native speakers from Riga. The sentences were recorded using a Sony WM-D6C recorder with a Sony ECM-121 stereo microphone on regular magnetic tape. The signal was digitized and analyzed using the Xwaves acoustic-phonetic analysis program on a Sun workstation. Segmental durations were recorded by measuring the signal in both the waveform and a wide band spectrogram. The measurements of the plosives include both the stop closure plus release burst, or measure from the end of voicing of the preceding vowel to the onset of voicing of the following vowel. The analyses were made using the statistical program S. All claims of statistical significance are based on t-tests set at the p<.05 level.

RESULTS

The experimental results confirm Laua's claims [4]. Both of my subjects showed the same patterns of segmental durations. Figures 1 and 2 show that (1) all voiceless consonants for both speakers are significantly longer than their voiced counterparts, (2) for the voiced consonants, duration is greater in the onset of the first syllable, and (3) for the voiceless consonants, duration is greater in the onset of the second syllable.

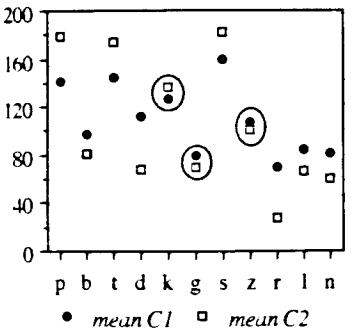


Figure 1. Mean duration (in ms) of consonants in onset of 1st syll. (C1) and onset of 2nd syll. (C2); speaker IL. Circled differences are not significant.

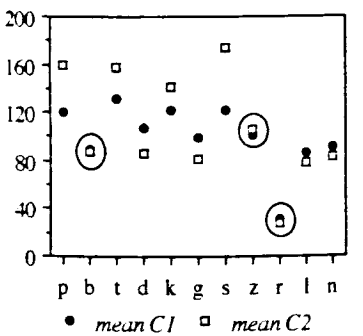


Figure 2. Mean duration (in ms) of consonants in onset of 1st syll. (C1) and onset of 2nd syll. (C2); speaker LL. Circled differences are not significant.

For simplicity of presentation, Figures 3 and 4 show only data from speaker IL. Speaker LL shows identical patterns.

Figure 3 illustrates that phonetic lengthening takes place only following stress. It is not dependent upon the position in the word alone. The word *nekad* 'never' is an exception with primary stress on the second syllable. In this figure, the mean duration of /k/ in *nekad* is significantly shorter than in all other positions.

Figure 4 shows that a voiceless consonant does not lengthen if preceded by a long vowel in the first syllable or if followed by a long vowel in the second syllable. All differences of mean duration between voiced and voiceless consonant pairs are significant.

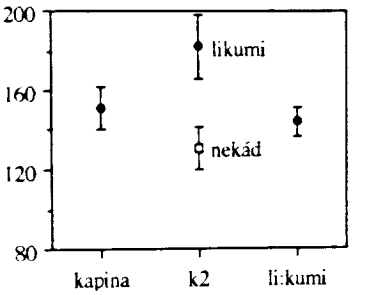


Figure 3. Mean duration (in ms) of /k/ in four words. Speaker IL. Error bars show standard deviation.

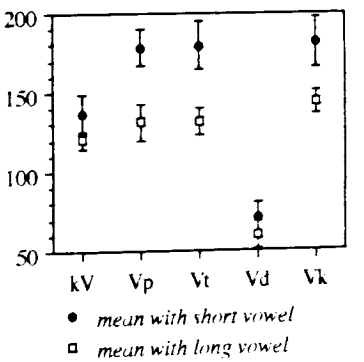


Figure 4. Mean duration (in ms) of various consonants in the onset of the 2nd syll. preceding and following long and short vowels. Speaker IL.

The explanation for the durational differences of /d/ in Figure 4 needs further investigation which is outside the scope of this paper.

Figure 5 shows that there is a significant difference for both speakers in the duration of /i/ in the third and fourth syllables of *nesalipinaːt* 'to not stick together', which could indicate a secondary stress on the third syllable [5].

Figure 6 shows that for the two word *nepametams* 'not discardable' and *nesalipinaːt*, the /p/ in the onset of the fourth syllable is longer in duration than in the onset of the second. This indicates a secondary stress on the third syllable precipitating the phonetic lengthening.

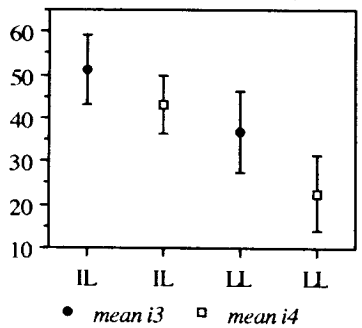


Figure 5. Mean duration (in ms) of /i/ for both speakers in nesalipina:t. *Third and fourth syllables.*

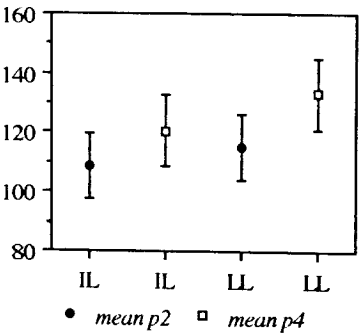


Figure 6 Mean duration (in ms) of /p/ in nepametams *(p2) and* nesalipina:t *(p4) for both speakers. All differences are significant at p<.05.*

I now turn to a phonological analysis to explain the distribution of the phonetic facts. The phonological analysis incorporates the new-found phonetic evidence for secondary stress.

PHONOLOGICAL ANALYSIS

In metrical phonology, word stress is associated with the presence of a metrical foot. Goldsmith [1] posits that Latvian words have a single unbounded left-headed foot, which is based upon his claim that Latvian words have only one stress per word, and that stress falls on the initial syllable. A problem with Goldsmith's analysis is that it does not account for the distribution of obstruent durations described above. However, the relationship between stress and consonant duration in Latvian can be understood when moras are taken into account [6, 7, 8]. I am here assuming that in Latvian, phonemically short vowels are dominated by one mora, while phonemically long vowels are dominated by two moras. In order to get the required stress patterns in Latvian, a simple Stress Condition needs to be posited for the language. *Stress Condition: Every stressed syllable must be heavy.* This means that every stressed syllable must have two moras. The first mora will by definition dominate a vowel, while the second mora can dominate either a vowel or a consonant. By assuming a quantity-insensitive system for Latvian, the first syllable will be subject to the Stress Condition, with the result that a second mora will be "inserted" in the first syllable even if that syllable has only a phonemically short vowel.

In addition to the Stress Condition, I am positing a Moraic Lengthening Rule (ML) shown in Figure 7. The second mora in the first syllable is "inserted" via the top-down assignment of metrical structure, as mentioned above.

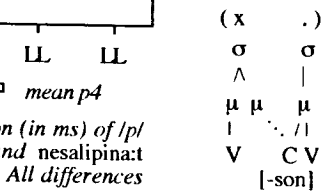


Figure 7 Moraic lengthening rule

This structure-building rule states that an obstruent in the onset of the second syllable of a foot will spread to fill an empty mora in the first syllable. The first mora in the first syllable can have an onset consonant or consonant cluster--this does not affect ML. In addition to ML, there needs to be a filter which does not allow a voiced obstruent to undergo the rule.

Applying this analysis to words in Latvian, a two-syllable word with a phonemically long vowel in the first syllable has the structure shown in Figure 9.

Figure 10 shows how the onset of the second syllable lengthens to fill the empty mora via ML if the vowel in the first syllable is phonemically short.

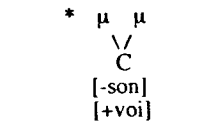


Figure 8 Filter

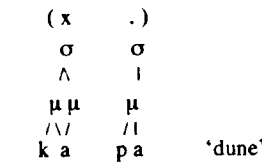


Figure 9. Phonological analysis of ka:pa.

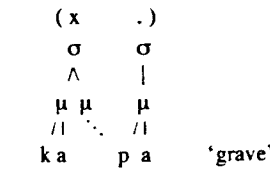


Figure 10. Phonological analysis of kapa.

In a word such as *vaga* 'furrow', the consonant will not undergo ML because of the Filter shown in Figure 8. The "empty" mora in the first syllable must either be deleted at the end of the lexicon, or else be unparsed in the output [9]. The /g/ in the second syllable cannot be syllabified in the first syllable, since syllables need onsets.

ITERATIVE BINARY FEET

Since there is phonetic evidence for vowel lengthening in the third syllable of a word, and for consonant lengthening in the onset of the fourth syllable, there must be a second foot with an empty mora in the third syllable able to undergo ML. Figure 11 shows that what Goldsmith [1] posits as the Foot layer in Latvian is actually the Word layer. Latvian metrical feet appear to be binary and iterative.

The methodology used in this paper cannot reveal the metrical status of a final phonemically long odd-numbered syllable. However, work by Hayes [6] and Kager [10, 11] suggests that such a final long syllable would indeed be footed, as shown in Figure 11.

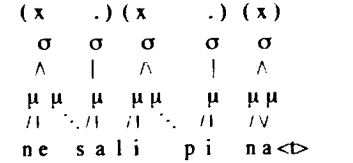


Figure 11. Possible phonological analysis of nesalipina:t.

CONCLUSION

The evidence provided in this paper suggests that Latvian has the following metrical system: a. Foot construction: i) Form syllabic (generalized? [6, 10, 11]) trochees from left to right, ii) Degenerate feet are not allowed. b. Word construction: End Rule Left.

Empirical research is still needed to determine whether Latvian builds syllabic or generalized trochees.

REFERENCES

[1] Goldsmith, J. A. (1990), *Autosegmental & Metrical Phonology*. Oxford: Basil Blackwell.

[2] Halle, M. and J.- R. Vergnaud. (1987), *An Essay on Stress*, Cambridge, MA: MIT press.

[3] Endzelins, J. (1922). *Lettische Grammatik*, Riga: Bildungsministerium.

[4] Laua, A. (1969), *Latviešu literārās valodas fonētika*, Riga: Zvaigzne.

[5] Bond, D. (1991), "Vowel and Word Durations in Latvian", *Journal of Baltic Studies* Vol. 22, 2, pp. 133-144.

[6] Hayes, B. (1993), *Metrical Stress Theory: Principles and Case Studies*, UCLA. Draft, January.

[7] McCarthy, J. J. and A. S. Prince, (1986), "Prosodic Morphology", Manuscript copy.

[8] McCarthy, J. J. and A. S. Prince, (1990). "Foot and Word in Prosodic Morphology: The Arabic Broken Plural", *NLLT* 8:209-283.

[9] Prince, A. and J. Smolensky, (1993), *Optimality Theory*, Unpublished manuscript, Rutgers University and The University of Colorado at Boulder.

[10] Kager, R. (1993), "Shapes of the generalized trochee", *Proceedings of the 11th West Coast Conference on Formal Linguistics*, pp. 298-312.

[11] Kager, R. (1992) "Are There Any Truly Quantity-Insensitive Systems?", *BLS*, 7:123-32.

NUCLEAR AND PRE-NUCLEAR TONAL INVENTORIES AND THE PHONOLOGY OF SPANISH DECLARATIVE INTONATION

Juan Manuel Sosa
Simon Fraser University, Vancouver, Canada

ABSTRACT

In this study I establish and describe the possible shapes, configurations and underlying tonal sequences of declarative utterances in Spanish. For this purpose I present the exhaustive inventory of tonal configurations that can occur in both the nuclear and prenuclear contours, -considered internal constituents of the intonational phrase-, and give an account of the actual occurrence and grammaticality of the possible combination of contours.

INTRODUCTION

Notation and Framework

The notation and abstract system of underlying intonationational units is based on the Pierrehumbert [1] theory of the phonology of English intonation. The contours of the intonational phrases are described as strings of L and H tones consisting of: (i) An initial boundary tone; (ii) A sequence of one or more pitch accents; (iii) A final boundary tone (H%, L%). The implementation of the different combinations of pitch accents and boundary tones determines the F0 contour.

Some Backround

It has been sustained by authors like Cunningham [2] and Kvavik [3] that that there is no fixed pattern of pre-nuclear intonation in Spanish. Or, that what they reflect is only sociolinguistic or expressive functions, not a real systematic structure.

Another common assertion about Spanish intonation is the idea, first expressed by Navarro [4], that the tonal peaks always correspond to stressed syllables.

The present research shows, on the contrary, that (1) There is a very definite and predictable pre-nuclear configuration for unmarked declaratives in Spanish; (2) The peaks actually correspond with the unstressed syllables following the stressed ones.

METHOD

Subjects and Data

For this research I analysed several hundred utterances from recorded spontaneous conversations of native speakers from both sexes and different ages, backrounds and national origin (from several countries in Central and South America, the Caribbean, Mexico and Spain). All individual utterances were analysed and classified according to their sentence-type, internal tonal constituency and semantic-pragmatic value.

Instrumentation

The acoustic analysis of the utterances was performed on the CSL 4300 of Kay Elemetrics. A number of altered pitch files (by means of the Paramenter Manipulation Option) was used to test the compatibility of some pre-nuclear shapes with the four characteristic terminal contours of declaratives.

NUCLEAR TONES

In previous work [5], I described the phonology and phonetics of all final contours (nuclear tones) of Spanish. The exhaustive list of possible declarative nuclei (combinations of final pitch accent and boundary tone) is the following:

L* L% (low fall); H+L* L% (low fall preceded by a high on the previous unaccented syllable); H* L% (high fall); and L+H* L% (high fall preceded by a low tone on the preceding unstressed syllable).

Not all of these, however, are characteristic of unmarked, neutral declarative utterances. By far the most common was he L* L% low-falling contour. In Figures 1 and 2 these final contours are illustrated.

PRE-NUCLEAR CONTOURS

Heads and Pre-heads

Following the British tradition, I refer to the pre-nuclear contours as 'head' and 'pre-head'. The head begins with the stressed syllable and ends with the syllable immediatly preceding the nucleus. The pre-head consists of any unstressed syllables before the first stressed one and since they are always low in Spanish declaratives, they are irrelevant for this analysis.

The head, on the other hand, is crucial to tune configuration and meaning, and is described according to their overall shape (level, rising, falling), as well as in their internal constituency in terms of pitch accents.

The "Bouncing Head"

The most common pre-nuclear pattern found for declaratives was overwhelmingly the progressively descending pattern. The pitch begins rather high, progressively stepping down with each peak until it reaches the tonal baseline with the final low fall L* L%. As can be seen in Figures 1, 2 and 3, the descent is not smooth; it takes place in a bouncy kind of way. As a rule, the stressed syllable is low, followed by a rise on the subsequent unstressed one, evidence of a sequence of recursive L*+H pitch accents. Ladd [6] and Gussenhoven [7] have characterized such multi-accented recursive patterns in heads as repetitions of the same contours.

On analogy with the names given by O'Connor and Arnold [8] to specific tone groups in a mnemonic, graphic way, I will call this pre-nuclear pattern the "bouncing head". Imagine a tennis ball bouncing two or more times on the court before rolling along the floor. The force comes from the bounce, which is the low-toned stressed syllable, followed by the rise on the following unstressed syllable, each successive bounce being lower. The following figures illustrate the most characteristic declarative utterances in Spanish:

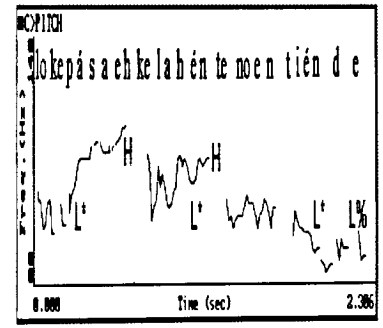


Figure 1. Typical unmarked declarative Spanish utterance with pre-nuclear L+H sequence and L* L% nucleus "Lo que pasa es que la gente no entiende" (male speaker from Honduras).*

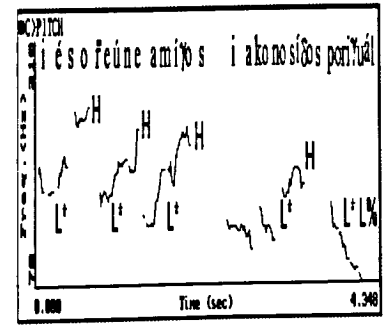


Figure 2. Uunmarked declarative Spanish utterance with four pre-nuclear L+H pitch accents and L* L% nucleus "Y eso reúne amigos y a conocidos por igual" (male speaker from Córdova, Argentina).*

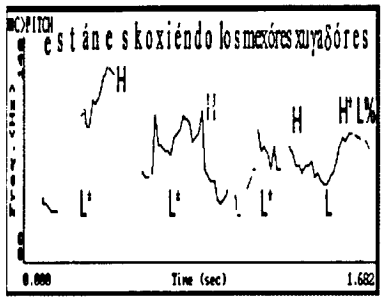


Figure 3. More emphatic declarative with pre-nuclear L+H sequence and L+H* L% nucleus "Están escogiendo los mejores jugadores" (male speaker from Guadalajara, Mexico).*

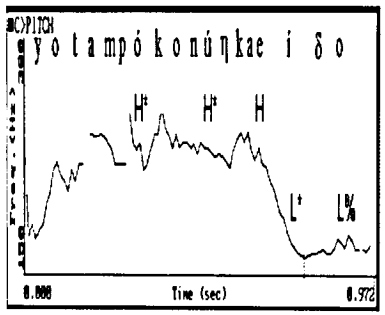


Figure 4. Categorical statement with high head and H+L L% nucleus "Yo tampoco nunca he ido" (female speaker from Lima, Peru).*

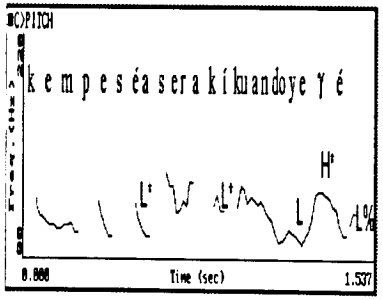


Figure 5. Declarative with low head and L+H L% nucleus "...Que empecé a hacer aquí cuando llegué" (male speaker from Medellín, Colombia).*

RESULTS

The Spanish Declarative 'tune'

The common overall pattern is a falling one, gradually descending into the final low fall. Although most descriptions agree on the descending "finality" of declaratives (Delattre, Olsen and Poenack [9]), to my knowledge only Bolinger [10] and Quilis [11] (implicitly), have described this pre-nuclear declining pattern. Most analyses have tended to repeat Navarro's [4] statement that the note of the first stressed syllable is more or less that on which the rest of the body of the unit is pronounced.

The canonical tune for unmarked declaratives in Spanish is then the combination of a bouncing head and a low fall. Or, in terms of L and H tones, a sequence of recursive L*+H pitch accents followed by a L* L% nucleus.

Variability and its Causes

There is of course a good degree of variability in the domain of declarative intonation, and that has been abundantly documented in the literature. However, it is not caused by factors such as age sex, country or region of origin or sociolinguistic status of the speakers. It rather seems to correlate with emotion, contrast and other kinds of emphatic speech. That is to say, utterances that may still be classified as declarative but have different configurations and tonal structure from the pattern here described, are instances of pragmatically distinct statements from the unmarked kind.

In our data we have statements with high head (Figure 4), low head (Figure 5) and also rising head, but they were perceived as different types of declarative utterances.

Just as Navarro [4] distinguished between "ordinary", "categorical", "dubitative" and "insinuative" statements according to the type of fall, it is possible to classify the pragmatic meaning of statements with different kinds of heads.

CONCLUSION

This research has uncovered an extraordinary agreement and regularity between dialects: the unmarked head for all dialects has the configuration that I have described as a combination of bouncing head (sequence of L*+H pitch accents), followed by a nuclear low fall L* L%. At least in their tonal underlying structure, there are no dialectal distinctions for this declarative sentence-type.

These findings contradict what has often been said, that there are no fixed patterns in Spanish intonational contours.

As I have shown, the peaks are always on unstressed syllables. At least for pre-nuclear contours, the common assertion that the highest point corresponds to stressed syllables is not in keeping with my findings. The stressed syllable is always low, followed by the high on the subsequent unstressed syllable, even across words and syntactic phrases. If there is at least one accented syllable in the pre-nuclear contour, the pitch accent will be L*+H.

To conclude, my findings show how remarkably predictable and regular the intonational structure of declarative sentence-types is. In the backround of the well-documented and striking dialectal and expressive variability of Spanish intonation, this is an significant finding. It shows that the intonational patterns are finite, systematic, characteristic and meaningful.

All these patterns and configurations can be economically generated by a model that includes pre-nuclear and nuclear contours as constituents, as well as the sequence of underlying tones that are the ultimate components of intonational phrases.

REFERENCES

[1] Pierrehumbert, J. (1980), *The phonology and phonetics of English intonation*, Cambridge, Ma: MIT Dissertation.

[2] Cunningham, U. (1983). "Aspects of the intonation of Spanish", *Nottingham Linguistic Circular*, 12, pp. 21-54.

[3] Kvavik, K. (1988). "Is there a Spanish Imperative intonation?", Hammond R and M. Resnick, eds, *Studies in Caribbean Spanish Dialectology*, Washington: Georgetown University Press, pp. 35-49.

[4] Navarro Tomás, T. (1944), *Manual de entonación española*, New York: Hispanic Institute on the United States.

[5] Sosa, J.M. (1991), *Fonética y fonología de la entonación del español hispanoamericano*, Amherst, Ma: UMASS Dissertation.

[6] Ladd, D. (1986). "Intonational phrasing: the case for recursive prosodic structure", Phonology Yearbook, 3, pp.311-340.

[7] Gussenhoven, C and T. Rietvelt. (1992), "A target-interpolation model for the intonation of Dutch", *ICSLP 92 Proceedings*, University of Alberta, pp. 1235-1238.

[8] O'Connor, J.D. and G.F. Arnold. (1973), *Intonation of colloquial English*, 2nd edition, London: Longmans.

[9] Delattre, P., C. Olsen and E. Poenack. (1962), "Acomparative study of declarative intonation in American English and Spanish", *Hispania*, XLV, pp. 233-241.

[10] Bolinger, D. (1961), "Three analogies", *Hispania*, XLIV, pp. 134-137.

[11] Quilis, A. (1981), *Fonética acústica de la lengua española*, Madrid: Gredos.

FORMAL AND FUNCTIONAL EVALUATION OF A MELODIC MODEL FOR STANDARD INDONESIAN

Ewald F. Ebing, Vincent J. van Heuven[1] & Cecilia Odé

Dept. Languages and Cultures of Southeast Asia and Oceania, Leiden University
[1]Dept. Linguistics/Phonetics Laboratory, Leiden University

ABSTRACT

A model of Indonesian intonation was perceptually evaluated using an improved testing methodology and listener selection. In a second experiment the focus and boundary marking functions of Indonesian intonation were investigated.

1. INTRODUCTION

A model for Standard Indonesian intonation has been developed following an analysis by synthesis methodology [1,2]. Successive versions of the model were perceptually evaluated by having native Indonesian listeners rate melodic versions of utterances (human originals versus model-generated contours, as well as *a priori* less adequate melodies, e.g time-shifted or Dutch contours) along a 10-point scale of formal melodic adequacy [3]. Listeners proved very insensitive to the melodic differences among the versions, so that we decided to re-run the evaluation with (hopefully) improved materials and more carefully selected listeners (section 2). It is difficult in Indonesian to distinguish between the accent-lending and boundary marking function of certain pitch movements [4]. In section 3, therefore, we examine how successfully Indonesian listeners can disambiguate arithmetic expressions with ambiguous focus distribution and internal bracketing.

2. FORMAL EVALUATION

Stimuli were taken from our corpus of quasi-spontaneous monologue by an educated speaker of Indonesian from Riau (East Sumatra) also used in our earlier experiments [2,3]. The stimuli comprised two tokens of the eight perceptually relevant pitch configurations found in our previous experiments. Four melodic versions of each configuration were produced by manipulating F_0 in the resynthesis (for procedural details see [3,6]:

a. *Close-copy* stylizations (COPY) of human originals; these should receive the highest ratings.
b. *Standardized* versions (STAN), i.e. generated according to our model; these should be (almost) as acceptable as COPY.
c. *Dutch*-based versions (DUTCH), generated according to the Dutch intonation grammar [1,3]; these versions should be rated as less acceptable than a or b.
d. *Mirrored* versions (MIRROR). Close-copies were mirrored along the frequency axis: rises became falls and vice versa; these versions should receive low ratings (as c).

The target configurations were now presented in their original contexts (rather than in isolation). To direct the listeners' attention to the relevant pitch configuration, the resynthesized context, but not the target configuration, was voiceless (whispered) throughout. This resulted in 64 stimulus types, each presented twice, yielding 128 judgments per listener.

The experiment was run at Universitas Islam Riau in Pekanbaru with 25 university students. Seventeen spoke Riau Malay as their first language, others had a different mother tongue, e.g. Minangkabau. Listeners rated each utterance along a 10-point scale of melodic adequacy (1: extremely poor; 10: excellent).

The results are summarized in Figure 1. The ordering of the acceptability ratings for the entire group of listeners is as predicted. No difference was found between COPY and STAN, t(783)=.149, ins., nor between STAN and DUTCH, t(777)=1.4, ins. However, the COPY versions were rated as significantly better than the DUTCH-versions, t(779)=3.12, p<.01. The MIRROR versions were rated as poorer than all other versions. Unexpectedly, STAN and DUTCH versions still do not differ significantly.

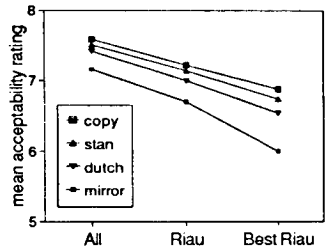


Figure 1. Acceptability of four melodic versions of Indonesian utterances broken down by listener selection.

We decided to enhance the effects by selecting only listeners with (i) the same variety of Indonesian as the speaker of the stimuli, and (ii) who were optimally sensitive to melodic differences.

First, the analysis was repeated for the 17 Riau listeners only. This time COPY and STAN versions do not differ from each other, t(538)=1.0, ins. but STAN and DUTCH do, t(532)=1.7, p<.05 (one-tailed). DUTCH and MIRROR versions differ as before, t(533)= 3.5, p<.01.

As the most sensitive listeners, only those eight Riau listeners were selected who obtained F>1 for the melodic version as a factor in listener-individual ANOVA's. Acceptability ratings are now better differentiated, while retaining the same ordering between conditions. These results show that the standardized pitch movements are perceptually adequate alternatives for close-copy stylizations. Moreover, to the Riau listeners, model-generated contours prove more acceptable than Dutch-based approximations. This confirms our hypothesis that the phonetic properties of the building blocks of Indonesian intonation are indeed language specific. Since the mirrored versions were included as a baseline condition, it is not surprising that they turn out to be the least acceptable. The fact that pitch contours that have been distorted in this manner are still rated in the upper half of the scale, is puzzling. Compared with results of similar experiments on English intonation [7], Indonesian listeners are remarkably tolerant towards deviations.

Finally, the difference between the whole group and the selected listeners suggests that regional and linguistic background does play a role: Riau listeners are more critical and discriminative.

3. ACCENTS AND BOUNDARIES

The aim of our second experiment was to find out to what extent accentuation and boundary marking can be (independently) expressed by means of the pitch movements in our model.

Focus distribution was manipulated by applying metalinguistic contrasts [5,6]. In the same set of test utterances, we also varied the position of a prosodic boundary by forcing the speaker to disambiguate a potentially ambiguous arithmetic expression (cf. e.g. [8]).

A single male native speaker of Indonesian produced eight versions of the same word sequence *dua kali tiga tambah lima*, orthogonally varying the position of the phrase boundary: *2x(3+5)* versus *(2x3)+5*, and focus structure:
(1) narrow focus on the first numeral
(2) narrow focus on the second numeral
(3) narrow focus on the third numeral
Each sentence was prompted by a question sentence to provide a context where one word was placed in focus. By manipulating F_0, model-generated contours were made for each realization.

The 25 subjects mentioned above indicated where they thought the speaker had intended the internal bracket of the expression to be, and - in a second part - which one of the three numerals in each phrase carried the strongest accent.

Table I specifies the percentage of accent responses for each of the three relevant numerals broken down by intended focus condition, and by intended phrase boundary position, first for the

model-generated pitch contours (A) and then for the human originals (B).

Table I. Perceived accents (%) for focus on 1st, 2nd and 3rd numeral, broken down by boundary position; (A) human originals, (B) model-generated contours.

A. Human	boundary after num. #1			num. #2			Δ due to boundary	
focus on num.	acc perceived on num. #1	#2	#3	#1	#2	#3	after #1	#2
#1	82	9	9	55	37	8	27	28
#2	16	80	4	6	93	1	10	13
#3	49	28	23	24	60	17	25	32
Mean	49	39	12	28	63	9	21	24

B. Model	boundary after num. #1			num. #2			Δ due to boundary	
focus on num.	acc perceived on num. #1	#2	#3	#1	#2	#3	after #1	#2
#1	45	25	35	48	40	12	-3	15
#2	18	54	28	19	62	19	-1	9
#3	25	22	53	16	43	41	9	21
Mean	29	34	37	27	49	24	2	15

In the human originals, accents on the first and second numerals are mostly correctly perceived, although the percentages are lower than we expected, and quite probably lower than what would be obtained with speakers and listeners of English or Dutch. Perception of an accent on the third numeral is strongly disfavoured. Crucially, there is a clear effect of the position of the internal boundary on accent perception: chances of perceiving an accent increase immediately before a phrase boundary. This effect is stronger when focus is on the first syllable than on the second.

For the model-generated contours, the same effects and interactions exist but in a weaker form. When the boundary is after the first numeral, the majority of accents is perceived on the syllables where they were generated, for all three positions: bias disfavouring the third numeral has disappeared. When the boundary is after the second numeral, some bias against perceiving accent on the third numeral remains, but it is clearly weaker than in the human originals. Apparently, our human speaker pronounced very clear accent-lending pitch movements on the first and second, but not on the third numeral. Our model-generated accents were identical for each numeral position, i.e. smaller than the human accents on the first two numerals, but larger than the human accent on the third numeral.

Again, there is an effect of boundary position on accent perception. This time, however, the effect is strongly asymmetrical: a boundary after the second numeral attracts many perceived accents onto the second numeral, but there is virtually no migration of accents to the first numeral when the boundary is after this numeral.

Table II specifies percent boundaries perceived after the first versus second numerals for the human originals (A) and the model-generated contours (B), broken down by intended phrase boundary position and intended focus condition.

Table II: Correctly perceived phrase boundaries (%) broken down by intended boundary position and focus distribution (A) human originals, (B) model-generated contours.

A. Human	boundary correctly perceived after		
focus on num.	numeral #1	numeral #2	Δ
#1	47	64	17
#2	27	79	52
#3	41	77	37
Mean	38	73	35

B. Model	boundary correctly perceived after		
focus on num.	numeral #1	numeral #2	Δ
#1	49	69	20
#2	34	66	32
#3	41	76	35
Mean	41	70	29

There is a very strong effect, both for human and for model-generated contours, for more (twice as many) boundaries to be perceived after the second numeral than after the first. It is unclear at this time to what extent this is a stimulus effect. A stimulus analysis (not presented) shows clear differences in duration structure as a function of intended boundary position, but the duration effects are in fact stronger for the first numeral than for the second. Therefore, it seems that the effect is due to linguistic expectancy.

There is a smaller effect, both in human and in model contours, to perceive

(10 percent) more boundaries after the first numeral when it is accented, and (10 percent) fewer when the accent is on the second numeral. In human contours there is a complementary effect to perceive fewer boundaries after the second numeral when the accent is on the first, and to perceive more boundaries after a second accented numeral; in the model contours, however, this interaction between accent and boundary position for the second numeral is no longer found.

From the above we conclude that the perception of accentuation and melodic boundary marking are intertwined. Boundaries are more likely to be perceived after accented words, and accents are more likely in pre-boundary position.

Identification of accents and prosodic phrase boundaries is only partly successful, both with human and model-generated pitch contours. However, asymmetries are stronger for the human originals. This may be due to the fact that the pitch movements used by the human speaker show large differences in excursion size as opposed to the standardized movements used in the model.

4. CONCLUSIONS

The formal evaluation of the proposed intonation model has shown that the pitch contours produced by the model are acceptable substitutes for (close-copy stylizations of) the originals. The functional evaluation allows to important conclusions to be drawn:

Firstly, is seems indeed true that the accent and boundary-marking functions are strongly intertwined in Indonesian; nevertheless, listeners were able, much better than at chance level, to distinguish between the functions. It is unclear at this moment whether this degree of interdependence is unusual. We know of no similar experiments, i.e. varying both focus and boundary positions, in other languages, so that we have no basis for comparison. Cross-linguistic experiments are essential for placing the performance of the Indonesian listeners, with both human and model-generated contours, in their proper perspective.

Secondly, formal evaluation of a melodic model (based on quality judgments) in itself is insufficient: it has to be complemented by a functional assessment of melodic adequacy.

ACKNOWLEDGMENT

Research supported by the Netherlands Organisation for Research through the Foundation for Language, Speech & Logic (project # 300-172-018).

5. REFERENCES

[1] Ebing, E.F. (1991). "A preliminary description of pitch accents in Bahasa Indonesia", in: *Proc. 12th Int. Con. Phon. Sc.*, Aix-en-Provence, pp. 258-261.

[2] Ebing, E.F. (1994). "Towards an inventory of perceptually relevant pitch movements for Indonesian", in: C. Odé and V.J. van Heuven (eds.), *Phonetic studies of Indonesian prosody*, Semaian 9, Vakgroep TC Zuidoost-Azië en Oceanië, RU Leiden, pp. 181-210.

[3] Hart, J. 't, R. Collier, A. Cohen (1990). *A perceptual study of intonation*, Cambridge University Press.

[4] Ebing, E.F. and Heuven, V.J. van (1994). "Some formal and functional aspects of Indonesian intonation", in: *Proc. 7th Int. Con. Austronesian Ling.*, Leiden (in press).

[5] Heuven, V.J. van, (1994a) "What is the smallest prosodic domain?", in: P. Keating, (ed.), *Papers in Laboratory Phonology III: phonological structure and phonetic form*, London (Cambridge University Press), pp. 76-98.

[6] Heuven, V.J. van, (1994b) "Introducing prosodic phonetics", in: C. Odé, V.J. van Heuven, eds. *Phonetic studies of Indonesian prosody*, Semaian, 9, Leiden (Vakgroep TC Zuidoost-Azië en Oceanië, RU Leiden), pp. 1-26.

[7] Pijper, J.-R. de (1983). *Modelling British English intonation*, Foris, Dordrecht.

[8] Lehiste, I., Olive, J.P. and Streeter, L.A. (1976). "Role of duration in disambiguation syntactically ambiguous sentences", *J. Acoust. Soc. Am.*, 60, pp. 1199-1202.

NATURAL EXPLANATIONS FOR PROSODIC CROSS-LANGUAGES SIMILARITIES

J. Vaissière
CNRS-URA 1027, Institut de Phonétique,
19 rue de Bernardins, 75005 Paris, France

ABSTRACT

A relatively wide range of crosslanguage prosodic similarities seems to be well explained if the typical prosodic configurations are assumed to result from a recursive implementation of similar patterns. Regardless of size, each contituent tend to conform the shape of a common archetypal (rise-fall) contour and a few derived contrastive (rise-non fall contours, whose characteristics may be motivated on biological, psychological and/or ethological grounds.

INTRODUCTION

A large of number of similarities in geographically and genetically unrelated languages may be explained if prosodic patterning is hypothesised to arise from a unique underlying archetypal fundamental frequency (Fo) pattern (cf. Lieberman's unmarked breath-group [1]).

I) Basic Rise-fall Pattern

a) Biological Base

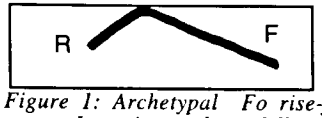


Figure 1: Archetypal Fo rise-fall pattern. Intensity tends to follow the same pattern.

The archetypal pattern is typically composed of an initial rise R followed by a fall F (Fig. 1). The continuous activity for obstructing the outgoing stream of air during expiration explains R and F as a departure from a rest position of the vocal folds. Long speech material exhibit a medial region where Fo raising and falling alternate with a slight general tendency for Fo to decline, the so-called declination line (DL) (Fig. 2a). None of the explanations for DL is completely satisfying, but DL is certainly bounded to physiological phenomena since DL tendency characterises infant cries [2] and some primate vocalisations [3].

b) Phonologisation of the archetype.

1) At the utterance level: The fall-rise pattern seems to have been "conventionalised" as the phonetic marker of the completeness of an utterance in many languages. R and F may be loosely bounded to the first initial and final word, respectively (Fig. 2a). Or R and F may be concentrated on a syllable in a particular position or to a lexically stressed syllable, depending on the language (Fig. 2b) [4] .

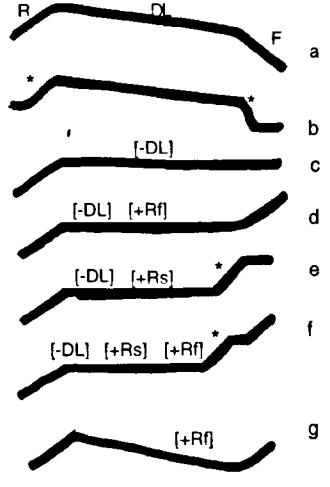


Figure 2: Fo in "complete" (a and b) and "uncompleted" utterances (c to g).
[-DL]: suppression of DL,
[+Rf]: rise on the final syllable
[+Rs]: rise on the last syllable.

2) At the word level:: A complete utterance can be composed by a single monosyllabic or multisyllabic word.

First, in tone languages, the rise-fall pattern may explain why there are more falling than rising tones(/F/ -> [falling tone] by default). Second, in languages with lexical stress, edge position, higher Fo and rising intensity tend to favour initial position for intensive or melodic stress. Edge position and "natural" final lengthening might also attract quantitative stress onto the final syllable. The low Fo and intensity, nasalisation and devoicing due to anticipation of the velic lowering and glottal opening required for breathing may however diminish the saliency of the final syllable in words spoken in isolation. Languages may also exploit a particular sensitivity of the ears to the duration of the penultimate syllable [5] (see Hyman's statistics on the location of word stress in 400 languages [6]): the initial position seems to be the most "natural" (stress marked by intensity and/or Fo?).

The archetypal pattern for the word as a statement in isolation is then more or less severely deformed in all its dimensions (Fo, intensity, duration), according to the position of the stressed syllable in the word and eventually the different type of accents (Swedish, Serbo-Croatian, etc.). There exists also a third category of languages without lexical stress. The pattern for French words spoken in isolation seems to be very close to the archetypal pattern. In connected speech, prominence perceived on the word initial or final syllable is due, grossly speaking, to focus and/or to syntactic boundaries.

The differences among languages and dialects can be explained (i) by the choice and the weighting of the acoustic correlates that signal word stress: Fo configurations, lengthening, higher intensity, spectral characteristics, (ii) by a different timing of the events relative to the stressed syllable and/or morpheme or word boundary [4]. The typical patterns found in isolated words isolation retains more or less of its duration, intensity and Fo characteristics when embedded in connected speech, and infants seem to be sensitive to the more frequent pattern of the maternal language [7].

II) NONFALL PATTERNS

a) Psychological and ethological explanations

The origin of the "natural" emergence of a non-falling pattern as a mark of uncompleteness may received either psychological or ethological explanations. First, according to Karcevskij [8], the (physical) lack of F is interpreted as a mark of lack of completeness. Second, for ethological reasons (Ohala's frequency code, [9]), low frequencies signal domination, so a person making a statement uses a low frequency,. High frequencies signal submissiveness. A person asking a question, in need of the goodwill of the hearer, tends to use a high pitch voice.

b) Phonologisation of the contrast

At higher level: contours typically observed in incomplete utterances, such as yes-no questions and non-final clauses typically do not end with a final fall. DL and F may be suppressed. An extra R may be eventually realized, bounded to the lexically stressed syllable of a focused word, or to the utterance final syllabe [+Rf] or penultimate syllable, and/or to the stressed syllable of the utterance final word [+Rs]) (see Fig. 2 c to Fig. 2g).

At the word level: Some languages like French, German, Spanish [10], Portuguese, and to a certain extend

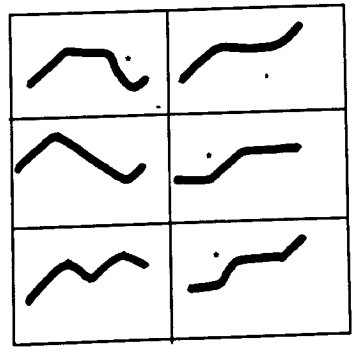


Figure 3: fall/ non-fall word contrasting patterns

Swedish (see Garding's work [11]), unlike English and Japanese, contrast strongly, in statements, between two accent patterns for words : a basic

falling pattern for words spoken at the end and a non falling or less falling (or even rising pattern) found for words in medial position in a sentence.

Rising word patterns may be used for the large majority of words not at the end of sentence, or at the edge of major boundaries only. The contrast generally affects the entire word. There is a striking similarities between word pattern and larger units patterns in a given language.

Division of a single sentence: A sentence often gives the impression of being composed of two parts: an overall nonfall portion and a falling portion [8] (fig 4). In neutral sentences, the "rising part" and the "falling part" generally characterise the thematic part (often the subject phrase) and the rhematic part of the sentences, respectively (cf. the use Garding's so-called grids [11].

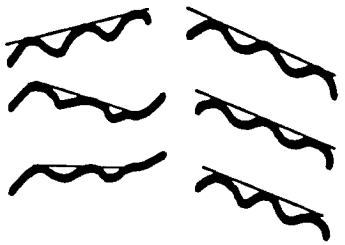


Figure 4: Illustrations of some of the various ways to divide the sentence into 2 parts: suppresion of DL, and/or adding of a rise in one the last syllables at the right edge of the non-fall portion.

Regrouping words: When two or more words are compounded into a single semantic block in a sentence, languages seem to maintain R in the initial word of each group and assign F later, giving to the block as a whole more or less the shape of a single word. This is interpreted again as a recursive implementation of the basic patterns.

III) FALL-RISE AS DISJUNCTURE

The Fo valley, which "naturally" occurs at the boundaries between sentences, seems to have been conventionalised as a boundary marker. An Fo valley superimposed on a long vowel may even create the impression of two vowels [12]. The boundary in pairs like "a name" and "an aim" is typically marked by a falling-rising pattern in the intensity curve [13] and Fo curve (glottal stop). The sharper the dip between the words, phrases, and sentences, the deeper the prosodic boundary [14]. The attenuation of the valley or its suppression (cf. the so-called hat-pattern), in contrast, expresses partial or total integration of the successive blocks. The use of a higher initial Fo value to mark the beginning of new informational units [15] and stronger fall and lower Fo values (such as at the end of a paragraph) to mark the end of a larger syntactic unit are commonly observed.

Downstepping of a High target preceded by a Low target can be interpreted as a phonologisation of the asymmetry in the production of the maximum speed of rises and falls inside a unit [illustrated in 16].

IV) REPETITION OF THE PATTERNS

Due to rhythmic effects and/or easiness in motor control, patterns tend to repeat or to alternate elements of similar size and shape: syllables, (strong, high, long and unreduced versus weak, low, short, and reduced), word and breath group patterns (rising and falling patterns ([17, 18]). The rhythmic tendencies are relatively independent of informational and syntactic characteristics of the sentences, and depends in part of the languages, the speaker, the rate of speech and the style.

CONCLUSION

The idea of an archetypal pattern is very much in line with Lieberman's unmarked breath-group [1]. The idea of superimposition of patterns of different sizes is in conformity with Garding's work on Swedish and Fujisaki's work on Japanese. What is proposed here is the recursive implementation of a few basic patterns, whose biologically, ethological and psychological bases lead to similarities among different languages.

Studies of more different languages are needed to confirm or disprove the psychological association of Fo rise and high values with the notion of beginning, Fo fall and low Fo values at offset, and of the fall-rise pattern as disjuncture. The question whether such psychological associations are valid also in tone languages is currently being investigated at the Institute of Phonetics in Paris.

Does it exist a number of basic prosodic principles which could predict the variety of prosodic systems attested and which would allow universal prosodic notation (Fo, duration, intensity and reduction characteristics) in terms of a reduced set of features? It is too early to answer such a question, because only a few acoustic studies are available, most of them dealing with only one prosodic parameter at a time. There are more phonological descriptions available but there show often divergent interpretations for the same language. Thanks to technological progress, it should be now feasible to construct and SHARE well documented data bases and international collaboration may help to overcome the unavoidable bias introduced by the maternal language of the investigators (and their former readings...).

ACKNOWLEDGEMENTS:

Thanks to my foreign students for collaboration, and also to N. Clements, J. Ohala and V. Kassevitch for lively discussions.

REFERENCES

[1] Lieberman, P. (1967). *Intonation, Perception and Language*. (MIT Press, Cambridge, Massachussetts).

[2] Lieberman, P.; Crelin, E. S.; Klatt, D. H. (1972). "Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man, and the chimpanzee". *American Anthropologist* 74: 287-307.

[3] Hauser, M. D.; Fowler, C. (1992). "Fundamental frequency declination is not unique to human speech: evidence from nonhuman primates". *JASA* 91(1): 363-369.

[4] Vaissière, J. (1983). "Language-Independent Prosodic Features. In Cutler, Ladd: *Prosody: Models and Measurements*,, (Springer-Verlag, Berlin), 53-66.

[5] Lehiste, I. (1979). "The perception of duration within sequences of four intervals". *J. Phon.* 7: 313-316.

[6] Hyman, L. (1977). "On the nature of linguistic stress". In L. Hyman, *Studies in stress and accent*, Univ. Of Southern California: Southern California Occasional Papers in Linguistics .

[7] Jusczyk, P.W., Cutler, A., and Redanz, N. (1993). "Preference for the predominant stress patterns of English words", *Child Development*, 64, 675-687.

[8] Karcevskij, S. (1931). "Sur la phonologie de la phrase". *Trav. Cercle Ling. Prague*, 4: 188-227.

[9] OHALA, J., (1994). "An ethological perspective on common cross-language utilization of Fo of voice", *Phonetica*, 41, 1-16.

[10] Delattre, P. (1962). "Comparing the prosodic features in English, German, Spanish, and French. *Int. Rev. Applied. Linguistics*. I: 193-210.

[11] G°arding, E. (1979). "Sentence intonation in Swedish. *Phonetica* 36: 207-215.

[12] Ainsworth, W. A. (1986). "Pitch Change as a Cue to Syllabification. *J. Phon.* 14-2: : 257-264 .[13] Lehiste, I. (1960). "An acoustic-phonetic study of internal open juncture". *Phonetica Suppl.* 5.

[14] Lea, W. N. (1972). "An approach to syntactic recognition without phonemics". *IEEE, AU*-21(3): 249-258.

[15] Lehiste, I. (1980). "Phonetic manifestation of syntactic structure in English". *Ann. Bull. RILP* 14: 1-27.

[16] Ohala, J.; Ewan, W. G. (1972). "Speed of pitch change". *JASA* 53: 345 (A).

[17] Vaissière, J. (1991). "Rhythm, accentuation and final lengthening in French". In Sundberg, Nord, Carlson, Eds.: *Music, Language, Speech and Brain*, pp. 108-120 (Macmillan Press, Houndsmills, England).

[18] Vaissière, J. (1994). " Caractérisation des variations individuelles du contour de la fréquence du fondamental observées dans des phrases lues en anglais", *20èmes Journées d'Etudes sur la Parole*, Société Française d'Acoustique, Trégastel.

LOW TONE VERSUS 'SAG' IN BARI ITALIAN INTONATION; A PERCEPTUAL EXPERIMENT

Martine Grice[1] and Michelina Savino[1,2]
[1]Institut fuer Phonetik, Universitaet des Saarlandes, Germany
[2]Istituto di Filosofia e Scienze del Linguaggio, Università di Bari, Italy

ABSTRACT

A perceptual test shows that the presence od a f0 dip just before the accented syllable plays a role in the perception of a yes-no question as opposed to a command in Bari Italian, and that this dip may be as small as 20 Hz. It is represented as L in a L+H* pitch accent.

INTRODUCTION

In Italian, both in the standard and in many regional varieties, including that of Bari, there are no syntactic or morphological means of distinguishing an information-seeking yes-no question from a non-question, such as a statement or command; instead, they are distinguished by means of intonation. For example, the sentence "Lo mandi a Massimiliano" can be interpreted as a question, loosely translated as "Will you send it to Massimiliano?", or a command "Send it to Massimiliano!" or statement "You send it to Massimiliano", depending on its intonation pattern. The actual pattern used differs from one local variety to another. Analysis of dialogue recordings involving six speakers of Bari Italian (B.I.) within a Map Task framework [1] has indicated that information-seeking questions (queries) have a rising-falling nuclear contour (see fig.1). Rising-falling contours also occur in questions of this type in Palermo Italian [2], [3], and in a number of other regional varieties of Italian [3]. We analyse the rising-falling pattern as a rising pitch accent on the focussed item, followed by a low target at the phrase boundary. The peak is around the middle of the accented syllable, and the preceding valley is on the preceding syllable. B.I. commands (instruct moves in [1]) have a falling nuclear contour where the fall begins around the beginning of the accented syllable (see fig.2). Within an autosegmental approach, we analyse the query contour as L+H* L-L% and the command contour as H* L-L%.

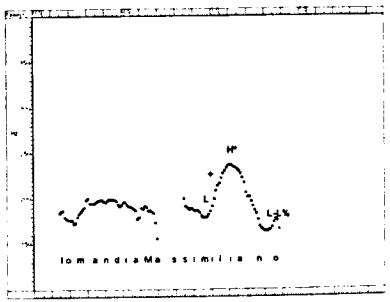


Figure 1. Natural rendition of query "Lo mandi a MassimiLIAno?"

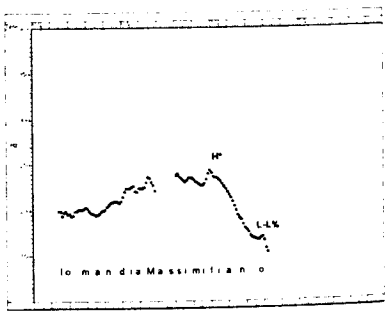


Figure 2. Natural rendition of command "Lo mandi a MassimiLIAno!"

Other autosegmental studies of Italian intonation, [2], [4], also use a L+H* pitch accent, but not in questions. Palermo Italian has L+H* L-L% in non-final items in contrast to L*+H L-L% in yes-no questions [2]; Standard Italian has L+H* in certain focussed items in non-questions [4]. Pitch accents with unstarred leading tones (i.e. T+T*) are more controversial in accounts of English (see [5] for discussion) than Italian. Not all autosegmental models of English intonation recognise L+H*; Ladd, for instance, treats L+H* as simply an emphatic version of H* [6]. One problem with the H* / L+H* distinction is that a H* pitch accent can be preceded by a degree of dip which is not attributed to a L tone, but rather to sagging interpolation. For example, between two H* tones which are far enough apart, Pierrehumbert's [7] model predicted sagging interpolation. This meant that a small dip would not be interpreted as a target L tone. Rather, it would be the result of an interpolation rule which applied automatically. A synthesis model [8] building on the foundation of [7] gave H* tones an underlying shape which, when filtered, created a peak accent with a small rise up to the local maximum. A L+H* pitch accent has a larger rise starting at a lower point.

Since it is clear that in B.I. the distinction between H* and L+H* is an important one, the aim of the perceptual experiment described in this paper is to investigate how low the local minimum has to be for a L tone to be perceived. The choice of sentence material precludes an analysis of the L tone as anything other than a leading unstarred tone of the nuclear pitch accent, since there is no potential prosodic boundary before the accented syllable 'LIA' of 'massimiLIAno', so the minimum cannot be attributed to a boundary tone of any kind. Perceptual experiment have been carried out on f0 peaks for German [9] and Hungarian [10]. Below it is the height of a f0 valley or dip rather than a peak which is under investigation.

THE EXPERIMENT

Stimuli

A number of renditions of the sentence "lo mandi a Massimiliano" were produced by a female Bari Italian native speaker, both as a query and as a command. These tokens were analysed, and a matching pair was selected where the tokens were closest in the f0 value of their endpoint and nuclear peak. F0 traces of the two tokens chosen are shown in figures 1 and 2. The f0 of each token was stylised using straight lines; and from the stylised tokens, two sets of stimuli were created using PSOLA resynthesis, as follows: Series Q: from the stylised query, 6 resynthesised versions were created as test stimuli by increasing (on a linear scale) the F0 value of the low target before the rise in 10 Hz steps up to the interpolation line between the two peaks; series C: from the original command, 6 resynthesised versions were created by lowering the F0 value at the same point in the preaccentual syllable corresponding to the low target in the original query in 10 Hz steps down to a position close (within 6 Hz) to the original query. A small difference in height of the two nuclear peaks (3Hz) and a difference in height and position of the first peak of the two natural stimuli (see figures 1 and 2) meant that the degree of dip from the interpolation line could not correspond to exactly the same Hz value at each step in the two series. The total set of test stimuli was 16 (2 original utterances + 2 series x 7 stimuli per series). The way the stimuli were constructed is shown in figure 3.

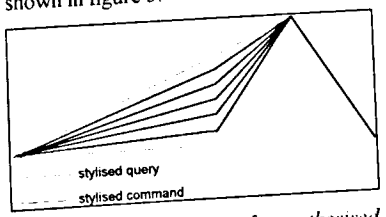


Figure3. Schematisation of resynthesised continua.

Subjects

Fourteen native Bari Italian speakers took part in the experiment. They were between 20 and 40 years old and were students and staff at the University and Politecnico of Bari. None of them had a background in phonetics.

Experimental procedure

The difference between the two communicative functions, query and command, was explained by means of examples in context. The 16 stimuli were presented five times, each in a randomised sequence in blocks of ten. Each stimulus was preceded by a 250 ms warning tone and 1 second silence. 5 seconds of silence followed for the subjects to respond. After each block of 10 stimuli there was a larger 11 second pause and a double tone of 250 ms as a precursor to the next block of stimuli. After each stimulus, subjects indicated on an answer sheet whether the utterance they heard was a query or a command. The total test duration was circa 20 minutes.

RESULTS AND DISCUSSION

Figure 4 shows the percentage of "query" responses for the series of stimuli with the stylised version of the query as the base stimulus (series Q) as a function of the size of the dip in f0 on the preaccentual syllable 'mi'. Figure 5 shows the percentage of "command" responses for the series originating from the stylised command (series C).

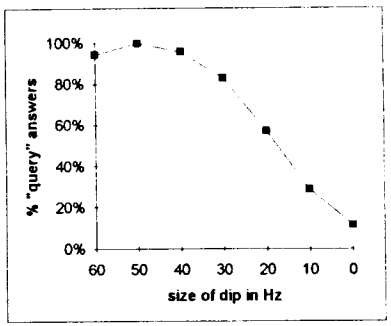


Figure 4. Percent "query" responses as a function of the size of the dip in Hz on the preaccentual syllable. Base stimulus originally query (series Q).

As it has to be expected, in set Q there is is a shift in response from query to command as the dip reduces in size, and in series C there is shift from command to query as the dip increases.

In series Q, the level at which more than 50% of responses were "query" was at a dip value of between 10 and 20 Hz (57% "query" responses for 20 Hz dip).

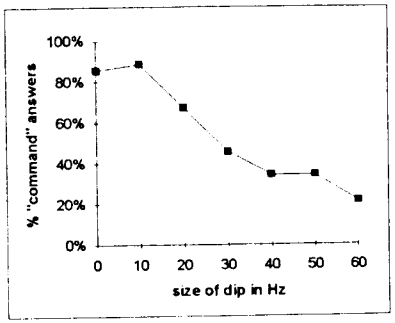


Figure 5. Percent "command" responses as a function of the size of the dip in Hz on the preaccentual syllable. Base stimulus originally command (series C).

This means that there only has to be a very small dip (considerably smaller that that observed in many cases of typically sagging interpolation, as exemplified in [7]) for a query and therefore L+H* to be perceived. However, in creating the stimuli from natural renditions, we neither manipulated nor held constant all the parameters which might serve to make the distinction between query and command. It could be, therefore, that there were other cues in series Q which led to the perception of a query, other than simply the degree of dip. It was with a view to controlling for some of these effects that a command was taken as the basis for the second series (series C). In this series, the degree of dip at which more than 50% of responses were "query" lay between 20 and 30 Hz (46% command, i.e. 54% query responses for 30 Hz dip). This means that the dip has to be lower in stimuli where the base token was a command. However, a 30 Hz dip is still comparable with cases in [7] of sagging interpolation in a similar pitch range.

It is apparent that the response scores in series C are less extreme than those in series Q. The base token in series C is recognised as a command in 86% of cases, whereas the base token for series Q is recognised as a query in 94%. The command which has been maximally altered (60 Hz dip) is only recognised as a query 79% of the time whereas the query with no dip at all is recognised as a command 89% of the time. Two factors may have influenced this.

1.Commands often display a larger downtrend in their baseline values than the one selected here for manipulation (other commands with greater downtrends had to be rejected because their endpoints were too low). The downtrend may be due to final lowering, as discussed in [11] in relation to Japanese, or to declination, as discussed in [12] in relation to Danish. In both Japanese and Danish, the downtrend does not occur in questions but does in other utterance types. The apparent flatter baseline could have made the stimuli in series C sound less command-like.

2.The presence or absence of a dip in the natural f0 contour appears to be linked to the position of the peak within the following accented syllable. After a dip, the peak is later (around the middle of the syllable) whereas without a dip, it is at the start of the vowel. Since the original peak position was retained in both series, it could have played a role in making the stimuli in series C sound less query-like, owing to the inevitably steeper rise from the dip. The cue for L+H* might not only be the presence of a dip in F0 before the accented syllable but also a later peak.

We can only speculate at this stage that such factors as final lowering (or declination) and peak and dip position may have made series C responses less differentiated. Further experimentation is needed to investigate these parameters.

Since we took two base tokens, one query and one command, and constructed two series of stimuli, we have confirmed that subjects can use a dip in f0 as at least one cue for discriminating between questions and commands. We have also provided support for the hypothesis that the dip in f0 constitues a target L tone, rather than an automatically sagging interpolation between two peaks, and shown that the dip does not have to be very low to be perceived as a L tone. This has implications for theories of intonation which allow for non-monotonic interpolation.

ACKNOWLEDGEMENT

We would like to thank Bill Barry and Mario Refice for advice and comments on an earlier draft of this paper.

REFERENCES

[1]Grice M., R. Benzmueller, M. Savino, B. Andreeva (this proc.), The intonation of queries and checks across languages: Data from Map Task dialogues.

[2]Grice M. (1992), *The intonation of interrogation in Palermo Italian: implications for intonation theory*, PhD dissertation, University College London, also (1995) Niemeyer L.A . series 334.

[3]Canepari L. (1992), *Manuale di pronuncia italian*, Bologna: Zanichelli.

[4]Avesani C. (1990), A contribution to the synthesis of Italian intonation, *ICSLP '90 Proceedings*.

[5]Grice M. (in press), Leading tones and downstep in English, *Phonology* 12.2.

[6]Ladd D. R. (1983), Phonological features of intonational peaks, *Language*.

[7]Pierrehumbert J. (1980), *The phonology and phonetics of English intonation*, PhD dissertation, MIT.

[8]Anderson et al. (1984), Synthesis by rule of English intonation patterns, *Proc. IEEE*

[9]Kohler K. (1987), Categorical pitch perception, *Proc. XI ICPhS*, Tallinn.

[10]Gosy M. and J. Terken (1994), Question marking in Hungarian: timing and height of pitch peaks, *J.Phon.* 22.

[11]Pierrehumbert J. and M. Beckman (1988), *Japanese Tone Structure*, Cambridge: MIT Press.

[12]Thorsen N. (1983), Two issues in the prosody of Standard Danish, in Cutler A. and D.R. Ladd (Eds.), *Prosody: Models and Measurements*, Berlin: Springer.

RHOTICS, JERS AND SCHWA IN THE HISTORY OF BULGARIAN

Georgi Jetchev
Scuola Normale Superiore, Pisa, Italy & University Sv. Kliment Ohridski, Sofia, Bulgaria

ABSTRACT

Three phonological contexts of early Slavic involving rhotics and jers, which gave different outcomes in the various Slavic languages, are explored. The Modern Bulgarian reflexes of these contexts are explained by a change in the way the listener corrected the acoustic signal: the tendency to over-correct it (which had given rise to syllabic rhotics) reversed into a tendency to under-correct it (which resulted in a later insertion of anaptyctic schwas next to rhotics).

DESCRIPTION OF THREE EARLY SLAVIC CONTEXTS

In the period of dialectal disintegration of Proto-Slavic, the two-level vowel system that characterized Early Slavic 2 [1, 2] was restructured in a four-level system through a shift from quantitative (fig.1) to qualitative contrasts (fig.2). The latter is the system used as the starting-point in the historical phonology of the individual Slavic languages. The mid vowels in it (levels 2 and 3) shared the feature [+lax]. The high lax vowels (ɪ and ʊ), traditionally called *jers*, demonstrated a tendency to reduction in some specific contexts: word-finally and before a syllable with a 'full' vowel, i.e. a vowel that is not a jer.

i i: ɨ: u u:

æ æ: ɑ ɑ:

Fig.1 Vowel system of Early Slavic 2

i ɨ u

ɪ ʊ

e o

æ ɑ

Fig.2 Vowel system of Early Slavic 3

During the Third Common Slavic Vowel Shift strong and weak jers developed very differently. Weak jers were lost whereas strong jers were retained as fully-fledged vowels and subjected to a lowering.

Context A

CirC, CurC (< *C$_r$C)

In Early Slavic 1 the syllabic rhotics of Proto-Indo-European developed leftward anaptyctic vowels (short *i* or *u*), thus becoming codas in rhymes with decreasing sonority: *r̥ > ir, ur [3:95]. Being in contradiction with the tendency only to admit rhymes with increasing sonority (the 'law of the open syllables'), the sequences of 'high vowel + rhotic' were most probably restructured once more in Early Slavic 2. In Old Church Slavonic we find the spellings "r + soft jer" (rɪ), "r + hard jer" (rʊ) as reflexes of Early Slavic 1 *ir, *ur.

Context B1

CrɪC, CrʊC before a syllable with a 'full' vowel and word-finally

This is the so-called weak position where jers were generally subject to loss. They are refferred to as *weak jers*.

Context B2

CrɪC, CrʊC before a syllable with another jer

This is the so-called strong position where jers were subject to lowering. They are referred to as *strong jers*.

PRESENTATION OF DATA FOR CONTEXTS A, B1 AND B2

Early Slavic 1 & 2 (reconstructed forms in IPA transcription):

A: /gurdla/ "throat", /virhu/ "top"

B1: /druva:/ "wood", /kristi:ti:/ "christen"

B2: /kruvi/ "blood", /kristu/ "cross"

Old Church Slavonic (attested written forms):

A: grʊlo, vrɪhʊ (грѫло, врьхъ)

B1: drʊva , krɪstiti (дрѣва, крьстити)

B2: krʊvɪ & krovɪ , krɪstʊ & krestʊ (кръвь & кровь, крьстъ & крестъ)

Russian

A: gorlo, verh (горло, верх)

B1: drova:, krestit' (дрова, крестить)

B2: krov', krest (кровь, крест)

Polish

A: gardlo, wierzch

B1: drwa, chrzcić

B2: krew, chrzest

Czech

A: hrdlo, vrch

B1: drva, křtiti

B2: krev, krest

Serbo-Croatian

A: grlo, vrh

B1: drva, krstiti

B2: krv, krst

Bulgarian

A: gărlo, vrăh (гърло, връх)

B1: dărva, krăstja (дърва, кръстя)

B2: krăv, krăst (кръв, кръст)

Czech merged contexts A and B1 developing syllabic /r/'s (see Table 1 where "V" stands for "vowel"). Serbo-Croatian merged all three contexts in a single reflex: syllabic rhotics.

In Bulgarian we find two different reflexes: ăr, ră. Moreover, there is a large set of Bulgarian words, historically related to contexts A, B1 and B2, which exhibit morphophonemic alternations (in either inflected or derived forms) with 'metathesis' of schwa (written ă) and *r* [4:166-200]. Here are some examples:

grăk 'Greek' ~ gărkăt 'the Greek', gărci 'Greeks'

vrăv 'twine' , vrăvta 'the twine' ~ vărvi 'twines'

krăv 'blood' ~ kărvav 'bloody',, okărvaven 'bloodstained'

Since 1899 Bulgarian orthography has been based on the following principle : "ră" is written (i) before 2 (or more) consonants; (ii) in monosyllables. In all other cases, i.e. before one consonant in polysyllables, "ăr" is written.

Compared to data from the other languages, the Bulgarian data suggest the following scenario: merger of A, B1 and B2 in a single reflex (to be found) and further differentiation in two different outcomes: ăr/ră.

Table 1. Reflexes of contexts A, B1 and B2 in the modern Slavic languages

	Context A *CurC, *CirC	**Context B1** CrʊC, CrɪC (weak jers)	**Context B2** CrʊC, CrɪC (strong jers)
Russian	V left to /r/	V right to /r/	
Polish	V left to /r/	no V	V right to /r/
Czech	no V		V right to /r/
Serbo-Croatian	no V		
Bulgarian	V either left or right to /r/		

ACOUSTIC CHARACTERISTICS OF RHOTICS IN SOME SLAVIC LANGUAGES

Bulgarian post- and pre-consonantal rhotics

Modern Bulgarian rhotics are apical taps and are typically realized as "an (almost) empty space on a spectrogram without any formants" [5:165-6], but only in intervocalic position. When they are preceded or followed by another consonant, a schwa-like vocoid element appears necessarily on oscillograms and spectrograms. In Bulgarian these svarabhakti vocoids (the term has been introduced by [6:298] in his description of Spanish rhotics) possess a formant structure very similar to that of a reduced vowel (schwa).The average duration of svarabhakti elements is about 30 ms.

Phonetically Bulgarian pre-consonantal rhotics represent a sequence of a tap and a svarabhakti vocoid (fig.3) whereas post-consonantal rhotics are a combination of a svarabhakti vocoid followed by a tap (fig.4). Compared to preceding (fig.3) or following schwa (fig.4), svarabhakti vocoids are shorter and of lower intensity.

Czech syllabic rhotics

The acoustic image of inter-consonantal rhotics in Czech is very similar to the

sequences "schwa + tap + svarabhakti vocoid" and "svarabhakti vocoid + tap + schwa" in Bulgarian. Czech syllabic rhotics represent a tap both preceded and followed by a svarabhakti vocoid (fig.5). The two svarabhakti vocoids are roughly of equal duration and intensity. Thus the acoustic image of Czech syllabic rhotics is symmetrical unlike that of the Bulgarian sequences "schwa + rhotic + consonant" or "consonant + rhotic + schwa", characterized by asymmetry.

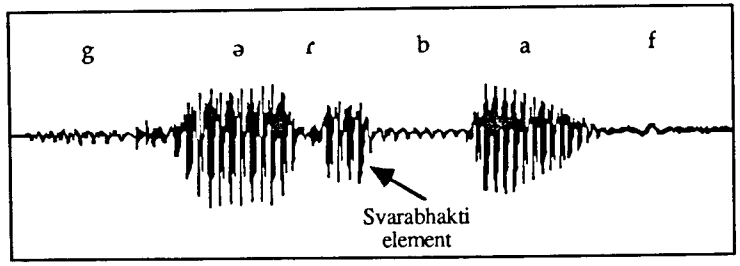


Figure 3. Oscillogram of Bulgarian pre-consonantal rhotic in gărbav *"hunchbacked"*

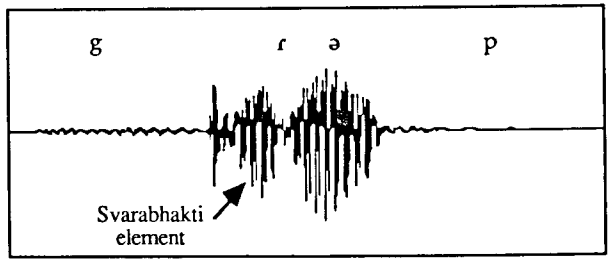


Figure 4. Oscillogram of Bulgarian pre-consonantal rhotic in grăb, *"back"*

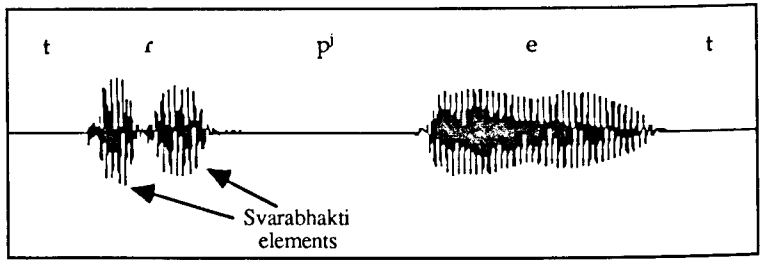


Figure 5. Oscillogram of Czech inter-consonantal syllabic rhotic in trpět, *"endure"*

SOUND CHANGES THROUGH UNDER- AND OVER-CORRECTION

As pointed out by Ohala [7:348] such "automatic" vocoids may create a sound change. The intervocalic rhotic is an unambiguous context and as such doesn't require any correction. There exists perfect correspondence between the speech signal and its perceptual interpretation. As for pre-consonantal and post-consonantal rhotics, they are contexts with virtual ambiguity: they require correction of the speech signal by the listener (factoring out of the svarabhakti vocoid).

Under-correction of svarabhakti vocoids. Anaptyxis.

If the listener fails to attribute the svarabhakti vocoid to the adjacent rhotic, he will misperceive it as a phonemic (most probably reduced) vowel. He will under-correct the signal. The resulting sound change will be an *anaptyxis*.

Over-correction of reduced vowels. Vowel loss.

If the listener inappropriately corrects the signal, he can misperceive a reduced vowel as a svarabhakti vocoid, erroneously attributing it to the adjacent rhotic. He will over-correct the signal. The resulting sound change will be a *vowel loss*.

A POSSIBLE SCENARIO FOR BULGARIAN

Merger of contexts A, B1 and B2

In context A, the jers in the Old Church Slavonic sequences ru, rı, lu, lı did not denote real jers. Even in strong position they were not subject to lowering. Words as vrihu, privu, skrubı never appear with *e, o* in the place of strong jers (i.e. in their first syllable).

In the manuscripts [8:139-140] there is often confusion between the hard and the soft jer in these sequences: pruvu instead of privu, srıdice instead of srudice, zruno instead of zrıno, etc.

By contrast, lowering of the jer in strong position does occur occasionally in the manuscripts where the sequences ru, rı, are found in context B2: krestu for kristu 'cross', krovı for kruvı 'blood'.

Hence Old Church Slavonic used the same spelling ru, rı (pъ, pь) for two different phonetic and phonological realities:

i. "svarabhakti vocoid + rhotic + svarabhakti vocoid" (phonetically), "syllabic rhotic" (phonologically);

ii. "svarabhakti vocoid + rhotic + reduced vowel (jer)" (phonetically), "nonsyllabic rhotic + reduced vowel" (phonologically).

The merger between contexts (i.) and (ii.) took place later: at the end of the Old Bulgarian period, that is, at the end of the 11th century.

The merger resulted from the reanalysis of the jer in context B as a svarabhakti vocoid, i.e. as part of a syllabic rhotic. This was a process of dephonologization. The perceptual mechanism which produced the sound change was that of over-correction. Listeners erroneously analyzed the jer as part of a syllabic rhotic. Then speakers began producing /r/ with symmetrically distributed svarabhakti elements (cf. fig.5) at the place of the earlier asymmetrical acoustic image corresponding to "rhotic + reduced vowel (jer)" (cf. fig.4).

Schwa anaptyxis

In a later period, a new tendency towards undercorrection of the acoustic signal arose. The Bulgarians then started perceiving either the leftward or the rightward svarabhakti vocoid of syllabic rhotics as a reduced vowel (schwa). This resulted in schwa anaptyxis. One of the svarabhakti vocoids was thus phonologized. Acoustically, this meant a return to asymmetry. The direction of the anaptyxis (leftward or rightward) depended upon the syllable structure. Apparently a constraint prohibiting "liquid + obstruent" codas was at work at that time. That is why "ăr" is admitted before one consonant when a vowel follows (i.e. in polysyllables), but not in monosyllables where the following consonant is word-final and hence it cannot be resyllabified as the onset of another syllable.

REFERENCES

[1] Andersen, H. (1985). "Protoslavic and Common Slavic – questions of periodization and terminology." *International Journal of Slavic Linguistics and Poetics*. 31-32: 67-82.

[2] Jetchev, G. (1994). "From early Slavic to modern Bulgarian: a survey of changes in the vowel system and the syllable structure." *Quaderni del Laboratorio di Linguistica, Scuola Normale Superiore di Pisa*. 8: 108-114.

[3] Carlton, T. R. (1991). *Introduction to the Phonological History of the Slavic Languages*. Slavica Publishers.

[4] Scatton, E. A. (1983). *Bulgarian Phonology*. Columbus, Oh., Michigan, Slavica.

[5] Lindau, M. (1985). "The Story of /r/." In V. A. Fromkin, ed., *Phonetic Linguistics*. Orlando & al., Academic Press. 157-168.

[6] Quilis, A. (1987). *Fonética acústica della lengua española*. Madrid, Gredos.

[7] Ohala, J. (1992). "What's cognitive, what's not, in sound change." *Lingua e stile*. 27: 321-362.

[8] Mirčev, K. (1958). *Istoričeska gramatika na balgarskija ezik*. Sofija, Nauka i izkustvo.

MODERN TENDENCIES IN STANDARD RUSSIAN SOUND SYSTEM OF THE END OF THE XX-th CENTURY

M. Kalentchouk
Moscow Pedagogical State University, Russia

ABSTRACT

The aim of present study includes working out a detailed description of the present-day younger generation pronunciational peculiarities; characterization of the development tendencies of Standard Russian sound system; revealing the factors which influence this development. The study is based on the results of the sociophonetic research of the younger generation speech.

There are two different types of sound laws. Some of them predict the realization of the phoneme in one possible way: synchronistically the word or the morpheme can be pronounced correctly only in a one sound form ([ва^{ъ}да] *'water'*, [друк] *'friend'*, etc.). The change of the pronunciation standard is connected with the phonetic regularities of a different type which describe the facts of synchronistic coexistence of pronunciation variants. For example, [в'э^{и}]*сна* and [в'и^{э}]*сна* *'spring'*, *було*[шн]*ая* and *було*[ч'н]*ая* *('bakery')*, *п*[о]*эт* and *п*[а^{ъ}]*эт* *'poet'*, etc. The variability of the pronuciation of the words is usually a stage in the transition from one nonvariable sound law to another. The present study is concentrated on the description of the sound variants.

Such variants may occur because of their belonging to different pronunciation sub-systems: sociolinguistic ones (chronological, local or sex) and language ones (connected with phonetic peculiarities of such groups of words as borrowings, proper names, functional words, terms, interjections, etc.).

Most of the available information about modern Russian pronunciation is based on the results of the investigations which were held in the middle of the century [1;.4;6]. A new generation has grown since that time and there is no doubt that their pronunciation differs from the one of the preceding generation.

In order to fix a "young" pronunciation standard the sociophonetic study of the speech of the younger generation was carried out [3]. It is based on the data obtained from different experiments: tape-recordings of specially composed texts read aloud (200 informants); responses to written questionnaires (1000 informants); the tape-recorded interviews (200 informants). Speakers were selected on the basis of demographic characteristics: all of them were Muscovites, their parents also came from Moscow; the sample includes young men and women (born in 1965-73), who studied in different Moscow universities. All speakers used the Standard form of Russian.

The comparison of two chronological subsystems - the present-day younger generation pronunciation peculiarities and the sound speech of the preceding generation constitute the major focus of this paper.

CONSONANTS

The main development tendency is the complication of consonant system, which is reflected in decrease of positional dependence of sounds [5, 336-343; 6, 442--446].

The process of overcoming softness of the sounds in the position before the soft consonant. In the beginning of the century the softness of the consonant before the soft sound was obligatory: *ю*[п'к']*и* *('skirts')*, *на ла*[м'п']*е* *('on the lamp')*, *че*[т'в']*ерг* *('Thursday')*, etc.

The studies of the speech of the middle of the century showed that consonant groups realized in three different ways: 1) for some sounds the old law is still working: dental+soft dental (*мо*[с'т']*ик* *'little bridge'*); [н] + [ч'], [ш:'] *бараба*[н'-ч']*ик* *'little drum'*, *же*[н'ш:']*ина* *'woman'*); 2) for some consonant groups a new law became obligatory: the sound is always hard before the soft consonant: labial + soft back lingual (*ла*[пк']*и* *'paws'*), etc.; 3) most of consonant groups allow variants in realization: dental + soft labial ([д'в']*ерь* and [дв']*ерь* *'door'*; labial+soft labial (*о бо*[м'б']*е* and *о бо*[мб']*е* *'about the bomb'*); dental or labial+[j] ([с'jэ]*хать* *'go down'*, [в'jу]*га* and [вjу]*га* *'snow storm'*) [1; 4; 5; 6].

The results of our experiments show that in younger generation standard overcoming of assimilation affected even such consonant groups which "opposed" this process longer than others (dental + dental). It is possible to assume that the process of overcoming assimilation has achieved a new stage. For different consonant groups this process started in a different time. What contributed to the unevenness of this process was the fact that some positions supported the pronunciation of a soft sound and some of them - the pronunciation of a hard one. The obtained data show that there is a contrast in behavior of consonants before the soft sound in a "yuong" standard inside the root and in sandhi (*пасьянс* *'patience'* and *съехать* *'to go down'*- the frequency of occurrence of the soft sound in the position before the soft is 100% and 38% correspondingly); in the frequently used words and in rare (*о банте* *'about the ribbon'* and *об обскуранте* *'about the obscurant'* - 95% and 55%); after the hard consonant and after the vowel (*шесть* *'six'* and *шерсть* *'the wool'* - 99% and 35%); inside the word and in the beginning of the word (*застеклить* *'to glaze'* and *стекло* *'glass'* - 94% and 68%), etc. According to orthoepic situation of the middle of the century it was possible to suggest that the next stage of the analyzed process will level the differences between various positions and it would lead to further overcoming. of assimilative softness. But regardless of expectations of linguists the process of assimilation is slowing down prolonging the stage of coexistence of soft/hard variants. The alternation of the consonant with the zero sound. The traditional norms of pronounciation of three and more consonants with [т, д] between the dental sounds or between the dental and [к] which demanded the alternation of the middle sound with zero sound are still working only for few con-sonant groups. The tendency for pronounciation of all sounds in such combinations affected more and more consonant groups (*стн, здн,стл, нтк, ндк, нтс, ндк*, etc). Such groups are pronounced with all sounds in actual phrase positions when the speakers want to emphasize some word; in terms; in rare words; in other cases they are used without middle sound. The alternation of the sounds [ж':] / [ж:]. Some Russian roots may be pronounced with the traditional variant [ж':] equally with modern [ж:] (*ви*[ж':]*ать* - *ви*[ж:]*ать*) *'to screech'*. The obtained data revealed that in "young" speech standard only 15 words are pronounced with the soft long consonant, most of these words are rather rare, and the word *дрожжи* *'yeast'* is the only one which is used with the soft sound more frequently than with the hard one. The alternation of the sounds [ш:'] / [ш'ч']. The correla-tion of these variants in younger generation standard is different in different positions. Inside the root only the sound [ш':] is used (*и*[ш':]*у* 'to look for'). But on the

morpheme boundaries both of the variants are used (ра[ш':]есать - ра[ш'ч']есать 'to comb'). First of all the choice of the variant depends on the type of morpheme boundary: on the boundaries of fusion tendency the sound [ш':] is more probable, but on the boundaries of agglutinative tendency the sound [ш'ч'] is used more frequently. The frequency of occurrence of sound [ш':] on different morphemic sature are following: preposition + word (40%); prefix + root (49%), root + suffix (83%)

VOWELS

The main development tendency is the simplification of vowel system, which is manifested in the increase of phoneme neutralization.

The modern model of phoneme neutralization in the unstressed position after the soft consonant is the following: 4 phonemes out of 5 which constitute the phonemic system of Russian vowels coincide in the sound [и^{э}] (<а> [ч'и^{э}]сы *'clock'*, <о> [в'и^{э}]сна *'spring'*, <э> с[т'и^{э}]на *'wall'*, <и> к[р'и^{э}]чать *'to shout'*) and only the phoneme <у> is realized in a different way ([л'у]бовь *'love'*). It's known that <у> is also involving in neutralization which leads to absolute simplificaion of the system [5, 22]. The obtained data show that according to the younger generation standard the pronunciations like [т'и^{э}]льпан instead of [т'у]льпан *'tulip'* are very rare. Such facts mostly occur in the words where there are conditions for assimilation with a vowel of another syllable: [б'и^{э}л'и^{э}]тень instead of [б'ул'и^{э}]тень *'bulletin'*. But there is one position - in the suffix of the Present Participle - where the variant [и^{э}] is pronounced even more frequently than [у]: ко[л'и^{э}]щий and ко[л'у]щий *'thrusting'*, подоба[и^{э}]щий and *подоба*[у]*щий 'proper'*, etc. In the unstressed positions after the hard consonant the neutra - lization model is more complicated - 4 different sounds are pronounced: [а^{ъ}], [ъ], [ы], [у] as a realization of 5 phonemes. In younger generation standard we see more evidence of the increase of neutralization: the phoneme <у> is being involved in neutralization after the hard consonants as well as after the soft ones (*г*[ъ]*бернатор* instead of *г*[у]*бернатор 'governor'*, *шт*[ъ]*катур* instead of *шт*[у]*катур 'plasterer'*); the replacement of unsressed [ы] by [ъ] in all positions except the 1-st pretonic and the final open syllables (in our experiments the auditors were unable to distinguish such pairs of words as домовой - *дымовой*, *выжить - выжать*, *тайнами-тайными*, etc., which can be an evidence of the fact that the speakers pronounced these words in the same way - *д*[ъ]*мовой*, *выж*[ъ]*ть*, *тайн*[ъ]*ми*).

The tendency for neuntralization is not working in marginal subsystems, which is the means of opposing them to general system: in borrowings (*п*[о]*эт 'poet'*), in terms *ф*[о]*нема 'phoneme'*), in interjections ([о]*го*), etc.

Other vowel variants which were analyzed in the present study - the sounds in the 1-st pretonic syllable after [ш], [ж], [ц] which correspond to letter *а*. At the beginning of the century the sound [ы^{э}] was pronounced in this position: ж[ы^{э}]*ра 'heat'*, ш[ы^{э}]ги *'steps'*. Nowadays the sound [ы^{э}] is replaced by [а^{ъ}] almost in all the words (ж[а^{ъ}]*ра*, ш[а^{ъ}]*ги*). But in some words the old variant is still in use (ж[ы^{э}]*леть 'to feel sorry'*, etc). The comparison of sound standard of two generations revealed that the number of words allowing such pronounciation is decreasing abruptly and we continue to use sound [ы^{э}] mostly in the words in which the sounds [и] or [э] are in stressed position what make possible the vowel assimilation (*двадц*[ы^{э}]*ти 'twenty'*, ж[ы^{э}]*смин 'jasmine'*, etc.

GRAMMAR FACTORS OF PHONETIC DEVELOPMENT

There are some grammar factors which can influence the sound system development: the tendency for agglutination in word-building and for analytism in morphology [2].

The tendency for agglutination is revealed first of all in the peculiar realization of phonemes in sandhi which makes the morpheme boundaries more obvious for speakers. The investigators of sound speech of the middle of the century determined the intensification of the juncture sygnals role: there was a contrast between a phoneme's realization inside the morphemes and on their boundaries (compare: *е*[с'л']*и* 'if'- [с#л']*ева 'to the left'*; [з'м']*ея 'a snake'* - ра[з#м']*енять 'to change'*; *бру*[ш:']*атый 'made of bars'*- [ш'#ч']*итывать 'to read from'* etc. But the results of the present study demonstrated different picture: in "young" pronuciation standard the signals of juncture are becoming less and less important due to the process of leveling the sound regularities inside the morpheme and in sandhi {compare: *е*[сл']*и* - [сл']*ева*; [зм']*ея* - *ра*[зм']*енять*; *бру*[ш:']*атый* - [ш:']*итывать*). It doesn't mean that there is no difference at all between the realization of phonemes inside the morpheme and in sandhi but the contrast is not so vivid as it was before.

Secondly, the intensification of agglutination in word-building can show itself in the tendency for uniform shape of the morpheme. For example, in "older" standard there was a pronunciation ра[з]*думать 'to change one's mind'*, but ра[з'jэ]*хаться 'to drive'*, *ле*[з]*ла 'she got into'*, but *ле*[з']*ли 'they got into'*, etc. In "young" standard the shape of morphemes in abovementioned words is not changing: ра[з]*думать* -ра[зjэ]-*хаться*, *ле*[з]*ла* -*ле*[з]*ли*.

The tendency for analytism in morphology is refleced in phonetic signals of word separateness. There is a phonetic feature which is able to manifest the grammar in dependency of a word or part of a word - the absence of vowel reduction in the unstressed positions. It may occur: 1) in prefixes and in first stems of shortenings, which aspire to have a status of a separate analytic word {*п*[о]*слезавтра 'the day after tomorrow'*; *д*[о]*обеденный 'before the dinner'*; *М*[о]*сбанк 'Moscow bank'*; [т'е]*хминимум 'minimum of technical knowledge'*, etc.) 2) in unstressed functional words (prepositions, conjunctions, particles) and pronouns (*вд*[о]*ль улицы 'along the street'*; *мороз*, *н*[о] *солнце 'it's sunny but cold'*; *м*[о]*й брат 'my brother'*. In all these cases the grammar independence of the language unit is weakened but such words and morphemes are "reminding" of their aspiration for "sovereignty" by a peculiar realization of vowels phonemes.

REFERENCES

[1] Avanesov R.I. (1984), *Russkoje literaturnoje proiznoshenije*, Moskva: Prosvetshenije.

[2] Glovinskaja M.J., Ilyina N.E. Kuzmina S.M., Panov M.V. (1971), "*O grammaticheskich factorach razvitia phoneticheskoy systemy sovremennogo russkogo jazyka*", Razvitije phonetyky sovremennogo russkogo jazuka, Moskva: Nauka, pp. 20-33.

[3] Kalentchouk M.L. (1993), *Orphoepicheskaja systema sovremennogo russkogo jazyka*: Dissertacia doktora philologichskich nauk, Moskva.

[4] Panov M.V. (1967), *Russkaja Phonetyka*, Moskva:Prosvetshenije.

[5] Panov M.V. (1990), *Istoria russkogo literaturnogo proiznoshenija XVIII-XX vv.*, Moskva: Nauka.

[6] *Phonetyka sovremennogo russkogo jazyka: Sociologo-lingvisticheskoje issledovanije* (1967), pod. red. Panova M.V., Moskva: Nauka.

STRONG AND WEAK CONSONANTS IN OLD AND MODERN GERMANIC

Anatoly Liberman
University of Minnesota, U.S.A.

ABSTRACT

As a phonetic parameter, consonant strength is inseparable from length, aspiration, and voice. Distinctive strength should not be recognized unless it is allowed to function word initially. In Germanic, only southern German has always fulfilled this condition.

In modern Germanic, the idea of consonant strength finds its main support in the functioning of obstruents in southern German dialects. A historian of Germanic observes strength in the study of the Second Consonant Shift, for intervocalic /p t k/ yielded /ff ʒʒ xx/; it appears that /p t k/ did not only change their place of articulation but were also reinforced. In word initial position, /p t k/ became affricates. To the extent that /pf ts kx/ are stops pronounced with a lax explosion, they can be looked upon as spliced and drawn out stops, i.e., as sounds homologous with /ff ʒʒ xx/. Intervocalic /p t k/ would probably also have become affricates, but they had to retain their independence vis-à-vis the reflexes of old /pp tt kk/, which yielded affricates and made /p t k/ seek new realizations. Later all the reflexes of the Second Consonant Shift in High German behaved as strong.

Notker's Law also testifies to the presence of the ancient correlation of strength, at least in part of Alemannic. According to this law, word initial /p t k/ occurred in Notker's dialect after a pause and after the nonsonorous final consonant of the preceding word, while sonorous word final consonants were followed by word initial /b d g/ of the next word. Since in Notker's system vowels and /l m n r/ were opposed to all obstruents, the main distinction must have been between sonorous and nonsonorous consonants. For Notker /p t k/ and /b d g/, along with the fricatives and affricates, were nonsonorous, i.e., voiceless rather than voiced, but his /p t k/ did not coalesce with /b d g/, as happened in Central German dialects. The feature distinguishing Notker's /p t k/ from /b d g/ was therefore the degree of sonority, even though it demarcated two classes of voiceless stops.

The greater the intensity of voiceless stops, the less sonorous they must be. Conversely, to remain voiceless and to acquire a measure of sonority, voiceless stops need a lax articulation. It is reasonable to assume that the nonsonorous voiceless stops were strong, whereas their less sonorous correlates were weak.

The history of /t/ can likewise be interpreted in terms of strength. By Notker's Law, /d/ < /þ/ (as in *daʒ* 'that') alternates with /t/, so Notker had /t/ that participated in the opposition of sonority, or strength (*taʒ* versus *daʒ*). But /t/ < /d/ (as in *tac* 'day') did not alternate with /d/. The sandhi phenomena subject to Notker's Law show that when stops were not affected by sonorous sounds, they were strong. The phoneme /t/ < /d/, strange as this conclusion may seem, was always strong.

Later events confirm this conclusion. In Middle High German (MHG), stressed syllables of disyllabic words were lengthened (either the vowel or the intervocalic consonant was affected), but /t/ remained short after a nonlengthened vowel. Apparently, short /t/ possessed the property (strength) that the other consonants acquired as the result of lengthening. MHG /m/ behaved in the same way, which comes as a surprise and makes it clear that our reconstruction is incomplete; sonorants could, most probably, also have been strong and weak, as is the case in many modern southern German dialects.

The existence of strong and weak consonants in modern High German dialects is an established fact, but strength is, as a general rule, synonymous with gemination: strong consonants are long, weak consonants are short. The question arises to what extent length is different from strength. Strong intervocalic consonants are always long, and the same dependence characterizes word final consonants in monosyllables. In High German dialects, a strong consonant tends to follow a short vowel, and a syllable containing a long vowel most often ends in a weak consonant. In nearly all dialects in which vowels were lengthened in monosyllables of the *Kopf* 'head', *Tisch* 'table', *Loch* 'hole' type, word final strong consonants underwent weakening.

The formula "short vowel + strong consonant versus long vowel + weak consonant", as it is known, for instance, in Middle Bavarian, is not in principle different from the formula "short vowel + long consonant versus long vowel + short consonant", as it is current in all the modern Scandinavian languages except Danish. Strength as a feature distinct from gemination should be posited only when it differentiates consonants in word initial position.

Previous discussion centered on High German, but the terms *fortes* and *lenes* are widely applied to all the other old and modern Germanic languages. However, outside High German only analogues of strong and weak consonants can be detected, and sometimes these analogues turn out to be false. For example, between the 13th and 16th centuries late consonant shifts took place in Germanic. They affected old obstruents in Icelandic, Faroese, and Danish and resulted in the dephonologization of voice and phonologization of aspiration in /p t k/:/b d g/. Although on a smaller scale, this process has also been recorded in Swedish, Norwegian, English, and Low German. Loss of distinctive voice could have been due to the new role of the syllable as the minimal unit of segmentation in later Germanic, but, whatever its causes, it did not make strength distinctive.

There is no gain in calling aspirated consonants in Icelandic, Faroese, and Danish strong, the more so because aspiration is rather a concomitant of *lenes* than of *fortes*, despite the widespread tradition to identify aspiration with strength. Nor will we learn anything new about Germanic if instead of describing /pp tt kk/, etc. as geminates we rename them *fortes*.

In many cases, voiceless consonants (given the correlation of voice) behave like *fortes*. Everywhere in Germanic lengthening in disyllables took place before voiced consonants more easily and earlier than before voiced ones. In High German, weak consonants were the "nonblockers" of lengthening, but elsewhere this function was performed by voiced obstruents.

As pointed out in connection with the history of MHG /t/, in words of the (C)VCV structure either the first vowel or the intervocalic consonant was lengthened. In the disyllables of Middle Danish, intervocalic /m/ prevented vowel lengthening, as it did in MHG: cf. Modern German *Hammer* 'hammer', *Sommer* 'summer' and Modern Danish *gammel* 'old', *komme* 'come'. In some dialects of Middle Swedish, /p t k/ blocked vowel lengthening (i.e., they resembled strong consonants), while in others they joined /b d g/ (and so resembled weak consonants). But Old Scandinavian had neither consonants like those which arose in High German by the Second Consonant Shift nor alternations of word initial obstruents of the Alemannic type (Notker's Law), and without them we lack the means for reconstructing an ancient correlation of consonant strength. The similarity between the role /m/ played in Danish and in German cannot be ascribed to chance, but more convincing arguments are needed to equate the consonant systems of Middle Danish and MHG with regard to strength.

In languages with the correlation of syllable cut (i.e., in all the West Germanic languages and Danish), analogues of the High German *fortes* and *lenes* exist too. When the contact is "tight" (*stark geschnitten*), or after a checked vowel, for example, in English *bid*, /d/ is phonetically stronger than /d/ under the "loose" accent (*schwach geschnittener Akzent*), or after a free vowel, for example, in *bead*. Even within one and the same prosodic type (*bid/bit*, *bead/beat*), vowels are longer before voiced than before voiceless consonants (the reason is the same: the relative weakness of voiced consonants), but these distinctions are not supported by the main feature that makes strength in High German an independent entity, i.e.,

by the alternation strong/weak in word initial position. Nor does consonant length go together with the correlation of syllable cut.

It appears that *fortes* and *lenes* in Modern Germanic exist only where they existed of old, i.e., in the southern dialects of German.

While studying consonant strength, we become aware of a paradox significant for phonology on the whole: it is sometimes easier to reconstruct past events than to analyze synchronic relations. Here are a few examples. In the opposition /b d g/:/p t k/, /b d g/ are marked if the distinctive feature is voice. Such is, for instance, the situation in Russian. Regardless of whether the word final obstruents of Modern Russian are identified with voiceless phonemes (in which case [prut] *prud* 'pond' or *prut* 'switch', sb., are phonemicized as /prut/) or assigned to different obstruents on morphological grounds (then *prut*, genitive *pruta*, is /prut/ and *prud*, genitive *pruda*, is /prud/), or called archiphonemes (then both are /pruT/) — all three solutions have been offered — the fact remains that in the position of non-discrimination only voiceless sounds are allowed to occur, so voice appears to be the marker of the opposition.

Despite the differences between the consonant systems of Russian and German, speakers of German will also agree that /b d g/ are marked and /p t k/ unmarked, for *Rad* 'wheel' (dative *Rade*) is related to *Rat* 'advice' (dative *Rate*) as *prud* (in Russian) is to *prut*. Even if we treat German /b d g/ as weak and /p t k/ as strong, /b d g/ will retain their status of marked members, however awkward it may be to call weakness marked when there is strength. On the other hand, in English, in which the opposition /b d g/:/p t k/ is not neutralized according to the German-Russian pattern, markedness and the nature of the marked feature are harder to define. Neutralization, unlike defective distribution, presupposes ambiguity: Russian [prut] is *prut* and *prud*, German [rat] is *Rat* and *Rad* (one of course looks for potential words, not for actual homonyms). Therefore, the non-occurrence of /b/:/p/, /d/:/t/, /g/:/k/ after /s/ (in whatever language) should not be confused with neutralization. In words like English *sketch* and German *Skizze*, *sk-* cannot be opposed to *sg-*, but neither form is ambiguous in the sense in which [rat] and [prut] are, so this case is different from the preceding one and sheds no light on the distinctive features of /b d g/ and /p t k/. And true neutralization of /b d g/:/p t k/ is lacking in English.

Although English /p t k/ are voiceless in comparison to /b d g/, aspiration is more important for their recognition. If the mark is tantamount to the presence of a feature, it is more natural to call English /p t k/ aspirated and marked. In Danish, Icelandic, and Faroese, in which voice plays an insignificant role in the production and perception of /b d g/, the situation is clearer than in English; hence the agreement among phonologists that here we deal with the marked (aspirated) /p t k/ and unmarked (nonaspirated) /b d g/. Swedish and Norwegian are close to English. There seems to be nothing wrong with recognizing voice as the distinctive feature of /b d g/ in these three languages (then /p t k/ will emerge unmarked), but it is equally plausible to treat /p t k/ as marked (aspirated). With regard to /b d g/:/p t k/, English, Swedish, and Norwegian are so different from Danish that it is preferable to set up models which will highlight rather than blur this difference.

Standard German also defies a unique solution: neutralization points in the direction of marked (voiced) /b d g/, while the factors that are valid for the analysis of English /p t k/ as marked (aspirated) are present here too. In southern German, consonant strength is indispensable for an adequate phonetic description, but the speakers' intuition and a consensus among scholars cannot replace a set of strict procedures. Such procedures (usually, neutralization) are not always available, and when they are, their results may be at variance with other, equally valid evidence.

It is curious that against such a nebulous background a historian of German easily discerns strengthening, for Old High German (OHG) *pf/ff*, *ts/zz*, *kh/hh* are obviously the reinforced variants of Common Germanic **p*, **t*, **k*. The replacement of distinctive voice by aspiration in Danish, Icelandic and Faroese is also easy to trace. Without this change the voiceless correlates of /l m n r/ in Icelandic and Faroese would not have arisen before old /p t k/. Nor would preaspiration have acquired its function of being the sole distinguishing element of forms like Icelandic *lappa* 'walk' and *labba* 'mend'.

According to universal belief, historical phonology is unable to overcome its limitation, namely, the disappearance of sounds whose properties it attempts to describe. This is indeed a severe limitation, but it is partly compensated for by the study of the process of change. Dynamics can often reveal the nature of oppositions better than the kaleidoscope of phonemes can do it. Phonemes in synchrony are not quite the same entities as phonemes in diachrony. This is why aphasia and the acquisition of speech by children lend themselves to phonological analysis exceptionally well. A changing phoneme is like a running person: both show the observer their otherwise latent features.

We can now return to Notker's Law, which is an especially characteristic example of the paradoxical interaction between synchrony and diachrony. Since in Notker's Alemannic dialect only word initial /p t k/ occurred after a pause and /b d g/ were disallowed, it follows that /p t k/, rather than /b d g/, were unmarked. This conclusion is borne out by the fact that, according to the rule of "consonant hardening" (*Verhärtung*), the same /p t k/ occurred in word final position, to the exclusion of /b d g/. (This rule characterized the entire area of High German.) In the opposition /b d g/:/p t k/, markedness belonged to greater sonority. Notker's Law can be reduced to the formula: sonorous after sonorous, nonsonorous after nonsonorous. Next to sonorous sounds (resonants and vowels), stops became quasi-sonorous as well. The active role of sonority also testifies to the markedness of /b d g/. It will be seen that the distinctive role of strength, with weakness being marked, has not emerged from this analysis (sonority sufficed to describe all the phenomena under investigation). Above, strength was tentatively deduced as the feature of /p t k/ from general phonetic considerations, but, in looking at subsequent lengthening, we immediately detect either strength or at least a feature of the same order.

Since the times of de Saussure linguists have prided themselves on differentiating between synchrony and diachrony. Roman Jakobson has gone a long way toward pointing up the dynamic nature of synchrony and the stable knots of diachrony. Our task consists not in wiping out the line between history and the present-day stage of language development: we should merely profit by certain tensions that exist between the two. In the days of descriptive linguistics, a great deal was said about the nonuniqueness of phonological solutions. The nonuniqueness principle is attractive in that it provides the researcher with a flexible model, but it also opens the door to all kinds of legerdemain. It seems that multiple solutions are the price we have to pay for the complexity of our material. Thus, German /b d g/ are voiced (and marked) from one point of view and nonaspirated (and unmarked) from another. In all the Germanic languages that underwent vowel lengthening in the structure (C)VCV, so also in Low German, the voiced intervocalic consonant behaved as though it were weak (see above), but in dialects with *Schärfung/Trägheitsakzent*, *stoottoon/sleeptoon* the distribution of accents in *and/ant* groups depends on the presence of voice in the obstruent, and in general nothing indicates that /p t k/ are strong in this area (Rhein-Limburg). Recognition of such contradictions is not a tribute to the hocus-pocus approach: God's truth need not flourish in a strait jacket.

Our material is often indeterminate, and we should use the results of phonetic change for retrospective solutions. For example, the strength of MHG /m t/ follows from lengthenings and gives credence to the idea of Notker's nonsonorous /p t k/ being strong. Synchrony and diachrony remain separate, but we no longer balk at interpreting the make-up of some phonemes in light of what became of them. Classical phonetics took this type of reasoning for granted. Modern linguistics will only gain if it shakes off part of the Saussurean dogma and uses common sense instead of structuralist rigor.

SONORITY AND THE H-SERIES IN GEORGIAN

Priscilla McCoy
Berkeley Speech Technologies, Inc., Berkeley, CA , USA

ABSTRACT

Georgian, a language in the Kartvelian or South Caucasian family of languages, possesses a complex verbal system. A feature of the Georgian verb is that it marks the subject, object, and indirect object of the sentence. There are two sets of indirect object markers; the set that occurs less frequently is referred to as the h-series [1]. This paper examines the positive correlation between acoustic features, sonority and the rules that govern the H-series.

INTRODUCTION

Georgian, a language in the Kartvelian or South Caucasian family of languages, possesses a complex verbal system. A feature of the Georgian verb is that it marks the subject, object, and indirect object of the sentence. Marking is by prefixes which occur before the root of the verb (or preradical vowel should there be one). There are two sets of indirect object markers; the set that occurs less frequently is referred to as the h-series in Aronson's Georgian: A Reading Grammar (1982) [1]. This name is derived from the third person marker of this series which is {h}. As I am interested in the third person markers in particular, this name becomes all the more descriptive. The reason for extracting these markers from their own overtly morphological sphere is that the h-series exhibits two interesting phenomena -- 1) the distribution of the h-series allomorphs illustrate an interesting parallel to a major division in acoustic theory; and 2) as the use of this prefix in Georgian is decreasing, the fact that course of its loss manifests a marked sonority hierarchy.

The H-series in the title is being used here as a cover term as the name provides a convenient designation for this prefix in its manisfestations diachronically as well as synchronically. Thus I am not interested in the h-series per se, but in a particular set of prefixes that occur both in the h-series and to a much lesser extent (or only diachronically) as subject markers.

This paper will be divided into three sections. First I will give an account of the H-series as currently prescribed in grammars (Vogt 1939) [9], then, from data in Shanidze (1980) [7], a brief diachronic perspective, with examples taken from other environments where these prefixes appear. Second, I will discuss some relevant elements of acoustic phonological theory as presented in Preliminaries to Speech Analysis (Jakobson, Fant, and Halle 1951) [4] and different views on sonority (Clements 1990) [2] as they relate to Georgian. Third, I will integrate the two sections and investigate what this does for the h markers and Georgian phonology. Third, I will integrate the two sections and investigate what this does for the {h} markers and Georgian phonology. This in turn may suggest some phonological priorities in Georgian. This paper looks at the h-series from the perspective of acoustic phonetics and examines the positive correlation between acoustic features, sonority and the rules that govern the H-series.

CURRENT USAGE

The h-series is the lesser used series of indirect object markers in Georgian. In grammars of current usage (Aronson [1], Dirr [3], Marr and Briere [5], Rudenko [6], Tschenkeli [8], Vogt [9]) the prefixes for the h-series are:

Indirect Object Markers in Modern Georgian

m -- first person
g -- second person
h/s/0 -- third person

The use of **h** , **s** , or zero is dependent on the following sound.

Distribution of h-series makers

h -- p, g , k , k', q'
s -- d , t , t , j , c , č , j', c', č'
0 -- elsewhere (all other consonants and vowels)

e.g. mo-**m**-cer-a	S/he wrote me
mo-g-cer-a	S/he wrote you
mi-**s**-cer-a	S/he wrote him/her/them

Written slightly differently the distribution for the third person markers might be:

H-series Marker Rule

h > h / __ p, g, k, k', q'
h > s / __ d, t, t', j, c, č, j', c', č
h > 0 / __elsewhere

At this point it becomes useful to examine this prefix diachronically to see its former full range of environments and to understand its current more limited ones.

The h prefixes/infixes also used to represent the second person subject marker. According to Shanidze [9] and others the **h** was derived from **x**. This can still be seen in two verbs in Georgian:

Remnants in the Second Person

x-ar second person, pres., 'to be'
mo-**x**-val second person, fut., 'to go/come'

Otherwise, the second person subject markers were h/s(s)/0. The distribution was as follows:

Distribution of {h} as Second Person Subject Marker

h > 0 /___ vowels
h > s /___ d, t, t, j, c, č
(s > š /___ j, c, č,)
h /___ elsewhere

This distribution of the **h** markers is considerably expanded. It includes all of the synchronic rule plus the remaining labials, liquids and nasals -- all of which are preceded by the marker **h**. Now it is possible to re-write the original rule thusly:

H-series Rule Re-written

h > 0 /___ vowels
h > s /___ T, C (T-dentals, C-palatals)
h /elsewhere (P-labials, K-velars, N-nasals, L-liquids)

This more expanded distribution was the same for the indirect object markers as well, that is what is called the **h**-series.

The result is a distribution that is uncomfortable in simple articulatory terms; **P** and **K** pattern together with **h**, and **T** and **C** pattern together with **s**. I will leave this problem for the moment and discuss some relevant acoustic features from Jakobson, Fant, and Halle [4].

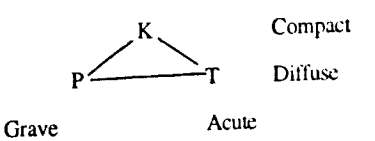
RESONANCE FEATURES

In Preliminaries to Speech Analysis [4], Resonance Features are introduced as a system that uses the acoustic signal to characterize divisions in the sound inventory. Resonance Features are then divided into I. basic resonator features: 1) compactness; 2) tonality features; and 3) tenseness and II. nasalization, using a supplementary resonator.

Consonants and vowels are divided into acoustic features as indicated by the patterning of their respective formants -- compact and diffuse. The features compact and diffuse are considered to be a primary split within the system. A secondary split, dividing consonants and vowels are the features grave and acute.

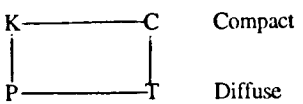
For languages such as French the consonants and vowels can each be set up on a triangle with for consonants a /t/ at the top and /p/ and /k/ at the bottom two points.

French



For a language such as Czech, the consonants (and vowels) are best divided on a square -- with the top corners being /t/ and /c/, form left to right, and the bottom corners /p/ and /k/.

Czech



In French, represented by a triangle, the secondary division of grave vs. acute cannot be manifested in both compact and diffuse consonants. However, in Czech, represented by a square, this is possible. Georgian patterns itself like Czech and uses a square representation of its consonants in this system.

Grave and acute, known as primary tonality features, pattern P, K together and T, C together. The second and third formants for P, K show similarities as do F2 and F3 of T and C. The formants in transition from vowel to consonant for the former are seen to be consistently upward moving, whereas those for the latter move downward into the consonant. Thus looking at acoustic features as opposed to articulatory ones, it is possible to find a natural patterning of P and K, which would be difficult in articulatory terms.

Going back to the rules for the distribution of the 3rd person markers of the h-series, it would seem that there is a positive correlation between the patterning of the features grave and acute, and the patterning of the h-series. Thus a single set of features seems to present itself as a possible explanation for the uncomfortable split if one opts for a solution in acoustic terms. As one can see, the features grave and acute comprise the linking of labials with velars, and dentals with palatals.

SONORITY

As a feature expressing sonority explains the distribution of the h-series, I will continue with the notion of sonority to explain as well the gradual loss of this prefix over time. Looking back in the second section, note the differences between the Old Georgian rule vs. the Modern Georgian rule.

Old Georgian	*Modern Georgian*
h>0 / __V	h>0 / __V
h>h / __ L, N	h>0 / __ L, N
h>h / __ P, K	h>h / __ P, K
h>s / __ T, C	h > s / __ T, C
s>š / __ C	

Sonority hierarchies are usually set up by the degrees of sonority, that is resonance. For Georgian, sonority becomes relevant for both the h-series morpheme and that of the following consonant. Clements [2] proposes a sonority hierarchy set up in such a manner.

+Sonority -Sonority

V > G > L > N > O

(V-vowel, G-glide, L-liquid, N-nasal, O-obstruent)

We assume principles of sonority operate in both the h-series and the followings consonant. The {h} becomes the target and the following consonant or vowel the trigger.

The sonority of [h] is closes to that of vowels, that is it is next in the sonority hierarchy after vowels. The first environment to lose {h} is pre-vocalic, vowels being the most sonorous. We therefore have two sonority hierarchies operating simultaneously -- that of the h-series where /h/ is more sonorous than /s/ and that of the following element. The friction of the laryngeal spirant [h] is produced when the air passes through the half-closed glottis. The noise then receives coloring from surrounding vowels. This may account for its earlier disappearance before vowels.

Next is {h} before L,N (Liquids and Nasals), most likely in that order:

\+ *Sonority* -

L > N

We now come to the remaining environments of the h-series -- obstruents. Because there has already been the change from h > s, we know that the next environment is most likely to have been that of the feature grave and finally the feature acute, with /s/ as its marker.

Both of these hierarchies of sonority, working simultaneously explain the loss of the h-series and the ordering of its loss.

CONCLUSION

In conclusion, the distribution of the h-series in Georgian (Old, Middle, and Modern) can be defined in acoustic terms using the feature grave/acute. Further the gradual loss follows a clear hierarchy of + sonority > - sonority .

As a final note, it is not clear how much the h-series in used in Standard Modern Georgian or if it is even present all the time in either Literary Georgian or Dialects.

[1] Aronson, Howard I. (1982), *Georgian : A Reference Grammar*, Columbus, Ohio: Slavica Publishers.
[2] Clements, G. N. (1990), "The role of the sonority cycle in core syllabification", in *Between the grammar and physics of speech*, ed. by John Kingston and Mary E. Beckman. Papers in Laboratory Phonology 1. Cambridge [England] ; New York : Cambridge University Press, 1990.
[3] Dirr, A. (1928), *Einfuehrung in das Studium der kaukasischen Sprachen*, Leipzig.
[4] Jakobson, R., G. Fant, and M. Halle (1951), *Preliminaries to Speech Analysis: the Distinctive Features and their Correlates*, Cambridge, Mass.: M.I.T. Press.
[5] Marr, R. and Briére (1931), "La langue Georgienne", Paris.
[6] Rudenko, B. (1940),*Grammatika gruzinskogo jazyka*, Moscow, Leningrad.
[7] Shanidze, Akaki (1980), *Txzulebani tormet tomad*, v.3, Tbilisi : Tbilisis saxelnicipo universiteti
[8] Tschenkeli, K. (1958), *Einfuhrung in die georgische Sprache*, Bd. 1, Zurich.
[9] Vogt, H. (1939), *Grammaire de la langue Georgienne*, Oslo.

Dari genesis: closer to Persian than to Tajik

V.B.Ivanov
Institute for Asian and African Studies
Moscow State University, Moscow, Russia

ABSTRACT

Spectral analysis of Dari vowels portrays long /ē/ and /ō/ more closed and higher than short /e/ and /o/. The situation was just the opposite thousand years ago in New Persian — the ancestor of contemporary Dari, Persian and Tajik. Contemporary Persian-Dari's /e/ and /o/ were short /i/ and /u/ at that time. Ancient /ē/ and /ō/ are pronounced now in Persian as long /ī/ and /ū/ resp. In the Persian-Dari's past short /e/ and /o/ interchanged their positions with long /ē/ and /ō/. Such a rearrangement did not occur in Tajik where short /i/ and /u/ united with their long neighbors /ī/ and /ū/. Two different processes (one common for Persian and Dari Vs the other one in Tajik) imply that New Persian was divided into two dialects: one belonging to Khorasan and the other one — to Maverannahr (two historical regions of Middle Asia). Lately Khorasan's dialect diverged into contemporary Persian and Dari while the Maverannahr's one became Tajik.

RELATIVITY OF PERSIAN, DARI AND TAJIK

New (Classical) Persian or *fārsi-ye dari* was a common language spread over the territory of contemporary Iran, Afghanistan and Middle Asia in XII — XV centuries. In the XVI century this linguistic community came to an end [1] and due to geopolitical reasons diverged into three closely related languages — Persian, Dari and Tajik. There is some evidence [2] that a certain difference in pronunciation appeared much earlier — in the XI century (/ā/ was pronounced like /ō/ in the Maverannahr region i.e. contemporary Tajikistan, Uzbekistan etc.). The sequence of appearance of these languages is not discussed in linguistic publications and a naive native speaker could think either they appeared all at once or still did not diverge at all, being 3 dialects of one language.

The vowel systems of contemporary Persian and Tajik were studied both articulately (by X-rays) and acoustically. But the positions of vowels in Dari vocalism were judged only by hearing. Some linguists suggest that long /ē/ and /ō/ (both called *majhul* "unknown" vowels, because they were not known to Arabs) are more open and lower than their short neighbors /e/ and /o/ [3], the other ones confirm just the opposite [4], [5]. An experimental study was necessary to make a well-founded conclusion that would help to compare the development of the three languages.

SPECTRAL ANALYSIS OF DARI VOWELS

The experimentation was based on a well-known concept that two first formants (F1 and F2) are related to the tongue position during vowel articulation. The same technique was used earlier to compare the properties of Russian and Persian vowels in bilingual pronunciation [6]. Despite common opinion that Persian /a/ is a front row vowel, our bilingual study proved it to belong to the middle row: Russian /a/ in *m'at'* "to crumple" (that was never considered to be a front vowel) is much closer to the front row than Persian /a/ in *madd* "tide".

Four Dari native speakers took part in a new experiment, the results of which can be seen in Table 1 and Figure 1.

Table 1. Formant frequencies of Dari vowels (in Hz)

vowels	F1	F2	F3
i	265	2125	3090
ē	400	2049	2820
e	420	1875	2675
a	695	1460	2475
ā	560	1085	2080
o	440	1025	1820
ō	410	905	1660
u	280	800	1485

In Table 1 and Figure 1 long /ē/ lies between /i/ and /e/. Similarly long /ō/ lies between /u/ and /o/. So both majhul vowels belong to the upper middle rise and are more narrow than their short neighbors /e/ and /o/.

Though it may seem that the F1 difference in pairs *ē/e* and *ō/o* is not much (only 20—30 Hz) the F1/F2-difference between them on the F1/F2 plane is significant: $p<0.1$ for *ē/e* and $p<0.03$ for *ō/o*. The difference in F3 testifies that the majhul vowels are more labialized than the short /e/ and /o/. Beside that the vowels differ in duration. Both parameters are significant ($p<0.03$ and $p<0.001$ — respectively).

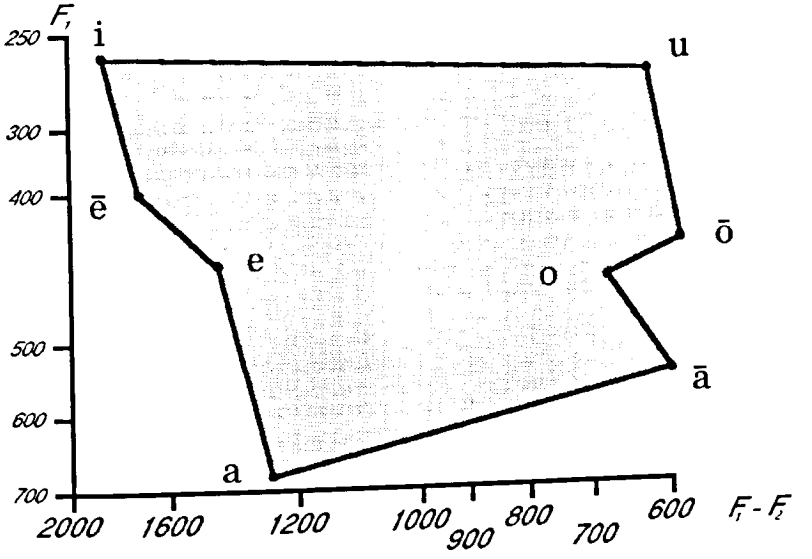


Figure 1. First and second formant positions of Dari vowels.

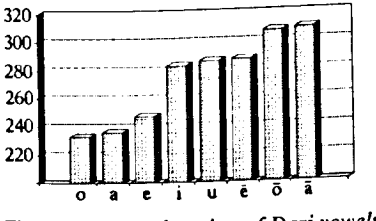


Figure 2. Mean duration of Dari vowels (in ms)

The mean duration, shown on Figure 2, divides the 8 Dari vowels into 3 classes: */e, a, o/* are short, */i, u/* — long, */ē, ā, ō/* — extra long. The extra long nature of */ā/* is caused by its openness: it is the most open vowel in Dari. The more open a vowel is the longer it sounds. Thus the extra long nature of */ā/* is a non-phonemic feature. But */ē/* and */ō/* are very closed and their extra long duration is phonemic. That's why */ē/* approximates the long vowels in duration.

Persian script used for official Dari writing does not show short vowels in most cases and does not distinguish *i/ē* and *u/ō* alternatives. It brings us to a unique situation in Dari not found in Tajik or Persian: Dari native speakers' identify vowels with difficulty. The vowels in triplets */e, i, ē/* and */o,*

u, ō/ can be interchanged depending on the speech style. In the official one the speakers try to use extra long vowels */ē, ō/* even if there is no historical ground for it, like in *arōs* "bride" (it is an Arabic word and must be free of majhul vowels). The same word can be pronounced *arus* in less official cases of literary language and *āros* in colloquial. The overall tendency in contemporary Dari is to substitute long and extra long literary vowels by corresponding short ones in colloquial speech: *sotun* > *soton* "column" , *budan* > *bodan* "to be", *nōzdah* > *nozda* "nineteen", *āwāz* > *awāz* "song", *āina* > *ayna* "mirror", *mēzanam* > *mezanom* "I strike" [7].

REARRANGEMENT OF EXTRA LONG AND SHORT VOWELS AFTER CLASSICAL PERIOD

In Ancient Persian there where 3 pairs of vowels. Inside each pair the vowels differed in phonological length: */ī, i/, /ū, u/* and */ā, a/* and 2 diphthongs */ai/* and */au/*. Those 2 diphthongs were the only diphthongs possible at that time: they were made by tongue movement from lower middle position towards extreme front or back. Such movements historically precede establishment of other diphthongs like */ui/* because the latter is formed across a catastrophic boundary which is a more complicated movement. Catastrophic diphthongs appear after the time the more probable non-catastrophic ones are already in use.

Later diphthongs */ai, au/* turned into monophthongs */ē, ō/* resp. [8] (Figure 3). */ī, i/* and */ū, u/* were articulated similarly, but */ā, a/* were different even then: */ā/* was closer to back row vs. more front */a/*. We can state it more precise that short */i/* and */u/* were non-significantly more centralized than their long counterparts, because generally it is difficult for the speaker to move the tongue during short period of time to an extreme front or back position. The centralization of */i, u/* allowed them later in Persian and Dari to reach the state of */e, o/* resp. It explains why it was just the short */i, u/* who did it but not the long neighboring vowels. The state of contemporary languages proves this hypotheses [9].

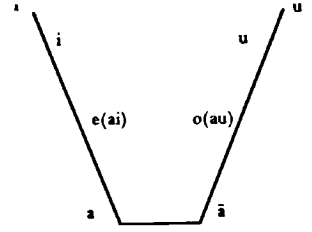


Figure 3. Vowel system of late Ancient, Middle and New Persian

After the classical period (about a thousand years ago) the previous system with 6 monophthongs and 2 diphthongs tended to be simplified. The diphthongs turned into extra long */ē, ō/*. It increased the number of rises: instead of one level of the middle rise two of them appeared. The Classical Persian began to branch.

In the Khorasan branch */ē, ō/* went to the upper part of the middle rise, in the Maverannahr branch — to the lower one. Later the Khorasan branch was divided into Persian and Dari. The situation in Dari which is well-known for its archaic elements remained just as it was in Khorasan dialect of the Classical Persian. Western dialects — Tehran's and Isfahan's Persian — developed further: narrow */ē, ō/* lost their phonological difference from */i, u/* resp. This process is still going on in Herat dialect which is an intermediate one between Persian and Dari [10].

Table 2. Front row vowels relativity. Underlined words contain vowels tending to higher rise

Persian	Dari	New Persian	Tajik	meaning
bist	*bist*	*bIst*	*bist*	twenty
bim	*bim*	*bIm*	*bim*	fear
xeš	*xeš*	*xiš*	*xiš*	brick
emruz	*emrōz*	*imrōz*	*imrûz*	today
miš	*mēš*	*mēš*	*mēš*	sheep
riš	*rēš*	*rēš*	*rēš*	wound

Table 3. Back row vowels relativity. Underlined words contain vowels tending to higher rise

Persian	Dari	New Persian	Tajik	meaning
dur	*dur*	*dúr*	*dur*	far
dud	*dud*	*dūd*	*dud*	smoke
sorx	*sorx*	*surx*	*surx*	red
xošk	*xošk*	*xušk*	*xušk*	dry
ruz	*rōz*	*rōz*	*rûz*	day
guš	*gōš*	*gōš*	*gûš*	ear

In the Maverannahr branch */ē, ō/* went to the lower part of the middle rise. It caused more narrow pronunciation of the short */i, u/* that finally in contemporary Tajik and Hazara dialect in Afghanistan merged with long */i, u/* resp. Tajik became a center row vowel, */ā/* went up to */ō/* causing former majhul */ō/* to be centralized */ù/*. In both branches the upper part of middle rise in back row was unstable and disappeared.

CONCLUSION

Some common features in Tajik and Dari like final */a/* that is not characteristic of Persian (Persian *xāne* ~ Dari *xāna* ~ Tajik *xona* "house") lead to the conclusion that the distance between Dari and Tajik was less than between Dari and Persian. But those differences and similarities (especially the tendency to pronounce */e/* instead of */a/* in Tehran and Isfahan in quite a number of positions) are product of later development. Global position-independent tendencies to mix up the majhul vowels with the long ones described above could not have been implanted into two neighboring languages by chance. Thus Dari and Persian should be considered closer relatives than Dari and Tajik.

REFERENCES

[1] Yefimov, V., Rastorgueva, V., Sharova E (1982). "Persidsky, tajiksky, dari", *Osnovy iranskogo yazykoznaniya, Novoiranskiye Yazyki*, Moscow: Nauka, p.7.

[2] Edelman, J. (1968), *Osnovnye voprosy lingvisticeskoy geografii (na materiale indoevropeyskikh yazykov)*, Moscow.

[3] Yefimov et al. Op. cit., p.26—27.

[4] Ostrovsky B. (1994) *Ucebnik yazyka dari. Cast 1.* Moscow, Nauka, p.19.

[5] Kiseleva L. (1985) *Yazyk dari Afganistana.* Moscow. pp.21—22.

[6] Ivanov, V. (1983) Russkiye i persidskiye glasnye v proiznoshenii bilingvov. Moscow, *Vestnik MGU, ser. Vostokovedeniye*, #2.

[7] Farhadi R. (1974) *Razgovorny farsi v Afganistane.* Moscow: Nauka, pp. 18—25.

[8] Edelman J. (1975) "Evolutia phonologiceskogo tipa". *Opyt istoriko-tipologiceskogo issledovaniya Iranskikh yazykov.* Moscow: Nauka, v.1, p.29

[9] Petito-Cocorda J. (1985) *Les catastrophes de la parole de Roman Jakobson a Rene Thom.* Paris: Maloine, p.283

[10] Ionesian Y. (1987) *Dialect sovremennogo Dari rayona g. Herata (phonetica, morphologia).* PhD Dissertation, Moscow.

THE ROLE OF PHONETICS IN THE EVALUATION OF RECONSTRUCTED SOUND CHANGE

J. Stuart-Smith,
Phonetics Laboratory, University of Oxford, Great Britain

ABSTRACT

This paper considers a typical case of reconstructed sound change whose mechanics are much debated. It is argued that an evaluation of the current explanations for the change is best made using an approach which begins with a consideration of the phonetics of the starting point, and then supports predictions for change with attested diachronic parallels. An evaluation thus made forces a rejection of the current theories and the proposal of a new explanation.

1. INTRODUCTION

A basic tenet of historical phonological reconstruction is that a phonetically plausible process of change connects the reconstructed starting point and the reflexes on which this is based. How the "phonetic plausibility" of reconstructed sound change is assessed seems to be rather vague. It is summarized critically by Lass [1, 171-2]: "our intuitive (or 'inductive') judgement of likelihood, based on pseudo-statistics of recurrence of change-types ... as well as (in some but not all cases) stipulations derived from knowledge of the kinds of articulatory or perceptual processes involved." Phonetic plausibility then seems to be assessed mainly by reference to intuitive feelings about how languages change.

In many cases, reference to parallel changes, and a general consideration of the phonetics of the sound in question do provide a fair guide. But there are some for which this approach is not useful; such a case will form the focus of this paper: the development of the Proto-Indo-European (PIE) voiced aspirates into the ancient Italic languages.

2. THE ITALIC DEVELOPMENT

"Italic" refers to a language group including Latin and her "sister" languages spoken on the Italian peninsula, attested from the seventh century BC (see Fig.1). The languages fall into two main groups, Latin/Faliscan, and Sabellian, to which Oscan and Umbrian belong.

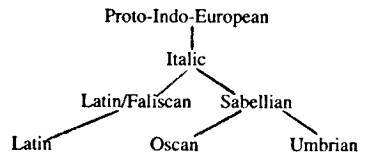


Figure 1. The Italic languages

The development may be summarized taking the labial stop as representative:

word-initial: PIE **bh-* > Lat. *f-*
Sab. *f-*

word-internal: PIE **-bh-* > Lat. *-b-*
Sab. -β-

The reflexes vary according to their position in the word (for the evidence, see, e.g. [2]). In word-initial position PIE **bh-* appears in Latin and Sabellian as *f-*, probably a voiceless labiodental fricative, thus PIE **bhrātēr* "brother" is found as Latin *frāter*, Oscan **fratrúm**. In word-internal position PIE **-bh-* appears as Latin *-b-* and Sabellian <f>, which represents a voiced fricative, probably bilabial or labiodental, given here as -β-. So PIE **tebhei* "to you" (dat. sg.) gives Latin *tibi*, Umbrian **tefe**.

The starting point of the change is usually taken to be a series of "voiced aspirates", phonetically breathy voiced stops, such as are found in contemporary North Indian languages, e.g. Hindi; their reconstruction for PIE is accepted here as unproblematic.

2.1. Previous Explanations

The changes have been explained by two competing theories, which I call here "Ascoli" [3] and "Rix" [4] (Fig. 2).

According to the "Ascoli" account, the PIE voiced aspirates devoiced in all positions in the word to voiceless aspirates, which then became voiceless fricatives. At this stage word-internal voicing occurred, leaving an allophonic distribution of voiceless fricatives in