COMMUNICATIONS
# PALATOGLOSSUS ACTIVITY DURING VCV UTTERANCES

## CONTAINING ORAL AND NASAL CONSONANTS OF HINDI

R. Prakash Dixit

Louisiana State University, Baton Rouge, LA, USA

## ABSTRACT

This study presents some EMG data from the palatoglossus (PG) and levator palatini (LP) muscles and examines the "gate-pull" model of active velar lowering for the nasal sound production.

### 1. INTRODUCTION

In January, 1972, Lubker et al. [5, p.235] proposed the "gate-pull" model of nasal sound production, which says that "...the levator may relax its activity in an almost gate-like fashion, thus allowing a temporal space during which palate is easily lowered. At some point in time during the "open" phase of the gate - or during the very early opening phase of it, a slight "pull" is provided by the palatoglossus to facilitate the ease and rapidity of palatal lowering. During this "gating" and "pulling" process the articulators function for the actual production of the nasal phoneme." However, various EMG studies of the PG muscle have produced conflicting results. The EMG data from PG reported by Lubker et al. [7, 8] on Swedish nasal consonants, by Fritzell [5] on English nasal consonants, by Benguerel et al [3] on French nasal consonants, and by Dexit at al [4] on Hindi front nasal vowels provided unequivocal support for the "gate-pull" model of nasal sound production. The PG data reported by Dexit at al [4] on back nasal vowels of Hindi were, however, primarily related to the tongue-body movement and positioning. On the other hand, the PG data on English nasal consonants reported by Bell-Berti [1], Bell-Berti and Hirose [2], and on French nasal consonants reported by Benguerel et al. [3] did not provide any support for the above model of nasal sounds production. Thus, the purpose of the present study was to explore whether the PG muscle is actively involved in lowering the velum for the production of nasal consonants of Hindi.

### 2. METHOD

Bipolar hooked-wire electrodes were used for EMG recordings. They were inserted peri-orally into the PG and LP muscles. (LP muscle data are a must for appropriate interpretation of PG muscle data.) EMG signals from these muscles were recorded simultaneously with audio signal while a native speaker of Hindi produced five repetitions of each: VCV nonsense utterances containing a nasal or an oral consonant. In these utterances, C represented /d/ or /n/, and V represented /a a u/. The first and second vowels in each utterance were the same, and the second vowel was stressed. EMG and audio signals were rectified, integrated and digitized. The offset of the first vowel was selected as the line-up point for ensemble averaging of the EMG and audio signals. Graphic illustrations of the ensemble-averaged EMG and audio signals were generated under computer control. They are presented in Fig 1.

### 3. RESULTS AND DISCUSSION

Figure 1 shows a high level of activity in the LP muscle for the utterances /iti/, /ata/, /udu/, /idi/, /ada/ and /udu/ containing an oral consonant. Whereas its activity is suppressed for the utterances /ini/, /ana/ and /unu/ which contain a nasal consonant, suggesting that the vowel's surrounding the nasal consonant in these utterances are fully nasalized. It is of interest to note that suppressed LP continues to maintain a certain level (though a low level) of activity in entirely nasal utterances, at least in this subject. Further, the consonant/vowel, and vowel height related differences (e.g., higher levels for consonants than vowels, and for high vowels than low vowels) generally observed in LP activity during oral utterances show up in nasal utterances as well. We will refer to these EMG patterns of the LP muscle in the description and discussion of PG muscle data below.

The PG muscle generally shows three peaks of EMG activity. The only exception is the utterance /ini/ where it shows only one peak. This lone peak in /ini/ and the last peak in all other utterances seem to be associated with velar lowering to open the nasal passage way at the end of the utterances, hence are of no concern to the topic of this study. Therefore, in this study, we will be concerned primarily with the presence or absence of the first two PG peaks. Incidentally, the PG muscle shows considerably higher peaks of EMG activity for the stressed (second) vowels as compared to those for the unstressed (first) vowels in Fig 1.

In this study, the PG muscle shows suppression of its activity throughout the utterance /ini/ which contains a nasal consonant surrounded by fully nasalized high front vowels. This suggests that PG is not involved in lowering the velum for the nasal consonant or the nasalized vowels in /ini/ and that the velum is lowered passively - simply by the suppression of LP activity for these nasal sounds. This finding for the Hindi nasal consonant is consistent with those reported by Bell-Berti [1], Bell-Berti and Hirose [2] on English nasals, and by Benguerel et al. [3] on French nasals, but inconsistent with those reported by Fritzell [5] on English nasals, and by Lubker et al [7, 8] on Swedish nasals. The finding on the front nasalized vowels of Hindi is rather unexpected, since in a previous study Dixit et al. [4] observed a high level of EMG activity in PG for the production of the front nasal vowels of Hindi. Similarly, in French a front nasal vowel was produced with a high level of activity in PG [3]. In these previous studies, however, the nasal vowels were contrastive, whereas in the present study they are contextually nasalized. Perhaps the PG muscle functions differently for contrastively nasal versus contextually nasalized vowels.

On the other hand, in the utterances /iti/ and /idi/ which contain an oral consonant in the front oral vowel context, PG shows two peaks of EMG activity. These peaks seem to represent its antagonistic or reflexive activity related to the tongue-body fronting by the genioglossus muscle for these high front vowels. Notice that LP in Figure 1 is highly active for the oral utterances /iti/ and /idi/ and suppressed for the nasal utterance /ini/. Thus the velum is in an elevated position for the former two utterances and depressed for the latter utterance. When the velum is depressed, the tongue-body fronting would not result in stretching the PG muscle, but when it is elevated, the tongue-body fronting would stretch the PG muscle, which may cause stretch reflex in this muscle. Lubker and May [9] have hypothesized such a stretch reflex in PG under similar physiological conditions.

In Figure 1, two peaks of PG activity are also observed for the utterances /ata/, /udu/ and /ada/ and /udu/ containing an oral consonant surrounded by the back oral vowels. Both these peaks appear to be
Fig. 1 Superimposed curves of ensemble averages of LP and PG EMG signals and audio signals for the experimental utterances. Audio and EMG signal amplitudes in arbitrary units and microvolts, respectively, are represented along the ordinate. The units along the abscissa represent 100 ms intervals. Zero (0) on the abscissa marks the line-up point used for ensemble averaging.

associated with the tongue-body movement and positioning for these back vowels. This is an expected result since LP shows a high level of activity throughout these utterances to stabilize the velum so that PG activity could contribute to the tongue-body movement and positioning (See condition or mode 1 in Lubker and May [5]). This result is in agreement with those reported in the other cited studies (particularly in [1,2,3,4]). In addition, two peaks of PG activity are also observed for the utterances /ana/ and /unu/ which contain a nasal consonant in the context of back vowels. Notice that LP activity is suppressed throughout these utterances as the back vowels surrounding /n/ are fully nasalized. However, EMG levels in the LP muscle for the utterances /ana/ and /unu/ never reach the zero level, that is the activity of LP is suppressed but not completely inhibited. As indicated earlier, LP maintains a certain level (though a low level, about 100 $\mu$V) of activity throughout nasal utterances. Because of this level of EMG activity in LP, it does not seem presumptuous to believe that the two peaks of EMG activity observed in PG for /ana/ and /unu/ are also related to the tongue-body movement and positioning for the back vowels surrounding the nasal consonant in these utterances. However, there is no PG peak that could be related to the nasal consonant in /ana/ and /unu/.

The above findings suggest that the activity of the PG muscle is primarily associated with the movement and positioning of the tongue-body for the production of oral and contextually nasalized back vowels, and antagonistically or reflexively related to the fronting of the tongue-body by the genioglossus muscle in the production of front oral vowels. The PG muscle does not appear to be involved in velar lowering either for the nasal or for the contextually nasalized vowels. Thus, the "gate-pull" model of nasal sound production fails to account for the results of the present study.

4. REFERENCES


ABSTRACT
Continuous speech of 23 subjects was recorded with and without masking noise. The group was composed of voice trained (n=12) and untrained (n=11) male and female Francophone subjects. The objective of the investigation was to find out how are spectral levels and voice quality affected under masked conditions for the different groups. Results show: voice trained subjects increase vocal levels less than untrained subjects under masked conditions, therefore showing an attenuation of Lombard effect. Some reported voice quality measurements (1:AB = 1000Hz / 1000Hz, 2.8:FO / FO) do not seem to apply to speech of Francophones.

INTRODUCTION
It is well known that the presence of noise produces an increase in vocal levels ([1]Lombard, 1911; [2]Lane and Tanel, 1971). Recently [4] Pick Jr. et al. (1989) suggested that through training the effect could either be enhanced or reduced but not completely eliminated. It is quite possible that people with voice training would be more apt to react differently to that effect. It has been shown, for example, that when singing in noise, trained performers deteriorate less than those of amateur singers ([6] Ward & Burns, 1978). That is attributed to a process of kinesesthetization, whereby vocal experience allows the performer to monitor the voice by proprioceptive rather than by auditory cues. Less dependent on auditory feedback, voice trained subjects would be less perturbed by noise and would therefore have the ability to preserve their voice quality. That ability should also be present in running speech. The objective of this study is to verify how are vocal levels of voice trained subjects affected when speaking in noise and whether voice quality is affected. The research questions are the following: 1. Are there long-term spectral level differences, at particular frequency intervals, of continuous speech, between voice trained and untrained subjects when speaking in noise? 2. Are there long-term voice quality differences, of continuous speech, between voice trained and untrained subjects when speaking in noise?
laterality effects. The recordings were performed at one foot distance from the microphone, and the order of the three conditions was systematically varied for succeeding subjects.

2.5. Analysis
The recorded samples were analyzed with an Ono Sokki CF300 spectral analyzer for Long Term Average Spectra at 1/3 octave intervals, 16-kHz range, for 128 spectra. The data was transferred and digitized in an IBM microcomputer through a software package designed for the project and then transferred to the mainframe computer where Spectral levels were determined for each of the three recording conditions.

3. RESULTS

The table above shows the following results: 1. There were no significant differences in the Normal Speech condition for spectral levels (F0e, F1e) at 3/10 octave intervals. 2. Spectral levels of voice trained subjects are significantly lower in both masked conditions (For SM: F0e, p<.01; F1e, p<.0005; BLK, p<.0005; ALK, p<.002) for SLH: F0e, p<.02; F1e, p<.002; BLK, p<.002; ALK, p<.04.

3. There are no significant voice quality differences (F1p, aAB) in the masked conditions between trained and untrained subjects.

4. DISCUSSION
There are no significant voice quality differences either in the normal nor in the masked speech conditions for the two groups. It is possible that the voice measurement method proposed for speech is linguistically related and therefore not appropriate for French. Trained Francophones do not have more energy in the region above 1000Hz relative to the lower frequencies. The other voice quality measures (S-SRM, S-LSM) were not significant, S-SRM, F1p, was proposed for singing. That might explain why it did not distinguish the speech of the voice trained. When speaking in noise, lower vocal levels clearly distinguished the voice trained from the voice untrained and confirmed that voice training diminishes the Lombard effect.

ACKNOWLEDGEMENTS

This work was supported by the Social Sciences and Humanities Research Council of Canada. The author wishes to thank Dr. Jean-Paul Dionne for his help and guidance in the statistical aspects of the project, and Michel Brabant for many hours of computer work in statistics.

5. REFERENCES

mean energy levels (dB) of voice THINH Francophones (N=12) and UNILAND Francophones (N=11) subjects for three speech production conditions measured over selected (1/3 octave) intervals.

<table>
<thead>
<tr>
<th>Speech condition</th>
<th>F0e</th>
<th>F1e</th>
<th>SRF0</th>
<th>BLK</th>
<th>ALK</th>
<th>aAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>21.73</td>
<td>22.14</td>
<td>-0.41</td>
<td>-17.93</td>
<td>-29.61</td>
<td>-11.67</td>
</tr>
<tr>
<td>Speech(S) Untr.</td>
<td>20.96</td>
<td>20.48</td>
<td>0.47</td>
<td>16.53</td>
<td>28.04</td>
<td>-11.51</td>
</tr>
<tr>
<td>Speech right T</td>
<td>21.06</td>
<td>19.86</td>
<td>1.20</td>
<td>16.60</td>
<td>26.64</td>
<td>-10.03</td>
</tr>
<tr>
<td>Speech left T</td>
<td>21.81</td>
<td>20.84</td>
<td>0.97</td>
<td>17.45</td>
<td>27.29</td>
<td>-9.83</td>
</tr>
<tr>
<td>Speech masked (S-RM)</td>
<td>18.54</td>
<td>16.71</td>
<td>2.14</td>
<td>13.62</td>
<td>23.60</td>
<td>-9.98</td>
</tr>
<tr>
<td>Speech masked (S-SLM)</td>
<td>0.66</td>
<td>-2.28</td>
<td>-1.62</td>
<td>-1.32</td>
<td>-2.97</td>
<td>-1.64</td>
</tr>
<tr>
<td>** significant at the 0.05 level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*** significant at the 0.01 level</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F0e: Energy at interval 80-160Hz for men, 160-250Hz for women
F1e: Energy at interval 315-600Hz
BLK: Energy below 800Hz (80-800Hz in 1/3 octaves)
ALK: Energy above 1000Hz (1000-5000Hz in 1/3 octaves)
ABSTRACT

Variations in the voice source for female speakers due to linguistic structure and speaker specificity have been investigated. The study is focused on consonants and transitional segments. The voice source has been analysed by inverse filtering. The consonant source spectra contained less energy in the higher frequency region compared to vowels. For a more leaky voice, transitional segments contained a larger amount of noise. Occurrences and origins of zeros in the spectra of voiced speech segments were studied using inverse filtering. For a leaky voice a zero was due to the incomplete glottal closure often occurred also in vowels.

1. INTRODUCTION

This study forms part of a project aimed at a complete description of female speech. The investigations have so far been on the male voice source. Information has been collected about the relationship between emphatic stress and voice source parameters [2] and about voice source variations with place of articulation of vowels [3]. The present study is focused on a description of consonants and transitions between voiced phonemes. Furthermore, the occurrence and origin of zeros in the spectra of voiced speech segments have been investigated.

The voice source was analysed by inverse filtering of the speech wave. A subsequent filtering of the LF voice source model [1] to the inverse filtered wave gave a parametric description of the voice source variations. The voice source parameters used in this study are RK, RG, EE, FA and FO. The source parameters are obtained from the quotient between the time from peak flow to excitation and the time from zero to peak flow. RG is the duration of the glottal cycle divided by twice the time from zero to peak flow. RG and RK influence the amplitudes of the lowest harmonics and are expressed in percent. EE is the excitation strength in dB and FA the frequency above which an extra -6dB per octave is added to the spectral tilt. In addition, the fundamental frequency, FO, is measured.

2. DYNAMIC VOICE SOURCE PARAMETER VARIATIONS

The present study concentrates on dynamic variations of the voice source. The rate of change of voice source parameters and how these changes correlate with segments and segment boundaries were investigated. For transitions between segments, especially between vowels and occlusive segments, both rate of change and the timing of changes are of crucial importance. For transitions between a vowel and an [i] or a nasal, the voice source parameter values change from typical vowel to consonant values within a few voice pulses. A transition between a vowel and [v] or [j] is much more gradual.

Correlations between the different voice source parameters have also been investigated. RG showed a fairly good correlation with FO in sentences uttered by different speakers. The correlation coefficient was found to be in the order of 0.75. Deviations occurred for FO peaks where RG was raised even more, see Figure 1. The remaining parameters did not show any substantial correlation with each other, the variations were more related to phoneme type and prosody. RK showed a large pulse-to-

VOICE SOURCES IN CONSONANTS

Voiced consonants in sentences have been inverse filtered, when possible, to achieve a source description. The investigated sentences contained the stops [b, d, g], the voiced fricatives [j, v] that both contain very little noise between voiced phonemes in Swedish, the nasal [n], the sonorants [r, l] and voiced [h]. It was often impossible to inverse filter the stops as they were too weak compared to the background noise. Accordingly, only one joint value was calculated for the stops. [r] was realized as a vowel-like segment in the studied sentences and had voice source characteristics similar to an unvoiced stop. To get a good fit between the LF voice source model and the inverse filtered wave-form for the remaining consonants, it was often necessary to cancel an extra pulse/zero pair, especially in [n], [l] and [v].

Voice source parameters for consonants and for some unvoiced vowels are given in the same sentences and given in Table 1. The values are averaged over at least 4 periods. Compared to vowels in the same sentence the consonants tend to have higher RK values, i.e. more energy in the lowest harmonics. The excitation amplitude, EE, was slightly lower for [r, l, n] than for vowels. For [v] and the stops, EE was often 10 dB weaker, the stops showed a rapid fall in EE through the sound. FA showed considerably lower values for all consonants with the exception of [r, l] and for one speaker [j]. A possible reason for the high FA is discussed below under "Noise excitation". FA was only slightly higher than FO for the remaining consonants. This means that the voice source contains less high frequency energy for these consonants than for vowels.

Table 1. Voice source parameters for voiced consonants and unvoiced vowels for two female speakers. The last column gives the number of occurrences of the phoneme in the investigated sentences. FO and FA are given in Hz and in percent. EE is given in uncalibrated dB so only comparisons within a speaker is possible.
4. VOICE SOURCE ZEROS

Zeros in voiced speech segments can have different origins. They are either a personal trait, often due to a leaky voice source, or a segment-related feature, especially in consonants, where it is due to the configuration of the vocal tract. Both these types of zeros have been investigated.

4.1 Zeros in consonants.

The investigated sentences contained consonants whose transfer functions contained zeros: [i] and [u]. For [i] and [u] the zeros and the connected pole are normally due to the geometry of the vocal tract. Zero/pole pairs found in [i] and [u] for two female speakers are given in Table 2.

The zeros sometimes detected in [v] as well as a low zero, about 900 Hz, sometimes found in [i], is presumably due to a more leaky voice source and consequently a coupling to the subglottal system in these consonants. This could be due either to an overall leaky voice or to a personal variation for these particular sounds. The zero/pole pairs are also listed in Table 2.

4.2 Voice source zeros in vowels.

Normally, while inverse filtering vowels, only anti-formant filters canceling the vocal tract resonances were used. For more leaky voices an additional pole/zero pair often had to be cancelled to achieve a good fit to the LF-model. The origin of this pole/zero pair is presumably a coupling to the subglottal system as for some consonants discussed above. The speakers who showed a zero/pole pair had a comparatively large amount of constant air flow during phonation in recordings with an Rotherenburg mask [5]. This implies an incomplete vocal cord closure and a coupling between the sub- and the supraglottal cavities. The frequency values of the pole/zero pair, a zero at about 800 Hz and a pole at about 1500 Hz, compares well with known values for subglottal poles and zeros for women [4]. In Figure 2 an example is shown of a vowel that has been inverse filtered using or not using an extra zero/pole pair.

5. NOISE EXCITATION

In inverse filtering and model fitting the model parameters tend to include the noise excitation since the inverse filter time window is one fundamental period. Accordingly, in a spectral section, no harmonics are visible and it is impossible to separate voice and noise excitation. This means that often a breathy segment will give quite high FA values contrary to theory. The high FA values for [b] and [j] in Table 1 are presumably due to this effect. To avoid this type of error, spectrograms of the utterances were studied. When a simultaneous voice and noise excitation could be suspected, partial inverse filtering was applied: all formants except one were damped out. The excitation pattern of the remaining formant showed if noise was a major excitation source. In Figure 3 an example of measured FA variations for a breathy voice and a more sonorant voice are shown. FA is highest during the transition from consonant to vowel for the breathy voice while FA is higher in the vowel for the more sonorant voice. The high FA values during the transition for the breathy voice turned out to be due to high noise content. We are presently trying to find a method to separate the two kinds of vocal tract excitations, this will be discussed further at the congress.

6. ACKNOWLEDGEMENTS

This project has been supported in part by grants from the Swedish Board for Technical Development (STU) and Swedish Telekom.

7. REFERENCES


Table 2. Zeros and corresponding poles in the voice source and vocal tract transfer function for some consonants. Zeros and poles are measured by inverse filtering. Zn denotes a zero frequency and Bzn its bandwidth. Pn denotes the corresponding pole and BPr its bandwidth. All are given in Hz. * denotes presumed voice source zeros and poles.

<table>
<thead>
<tr>
<th>Z</th>
<th>B1</th>
<th>Z2</th>
<th>B2</th>
<th>P1</th>
<th>P1</th>
<th>P2</th>
<th>P2</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>940</td>
<td>300</td>
<td>2050</td>
<td>300</td>
<td>1450</td>
<td>150</td>
<td>2500</td>
</tr>
<tr>
<td>W2</td>
<td>990</td>
<td>450</td>
<td>2200</td>
<td>300</td>
<td>1600</td>
<td>200</td>
<td>2550</td>
</tr>
<tr>
<td>W1</td>
<td>750</td>
<td>450</td>
<td>1900</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W2</td>
<td>860</td>
<td>150</td>
<td>1700</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W1</td>
<td>700</td>
<td>250</td>
<td>1650</td>
<td>150</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W2</td>
<td>840</td>
<td>350</td>
<td>1600</td>
<td>200</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W1</td>
<td>1850</td>
<td>600</td>
<td>3000</td>
<td>500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W2</td>
<td>2300</td>
<td>600</td>
<td>2600</td>
<td>450</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. FA variations in a transition from [j] to [e]. F2 is plotted to illustrate the transition. The left half shows a leaky voice, the right part a more sonorant voice.
TEMPORAL MODELLING OF GESTURES IN ARTICULATORY ASSIMILATION

Martin C. Barry

Department of Linguistics, University of Manchester, England

ABSTRACT

Gestural trajectories for consonants in coronal + velar clusters were derived using EPG contact data from speakers of English and Russian. Evidence from rapid speech indicates a variety of articulatory strategies available to speakers of the two languages, with notably a high-level discrete assimilation process found only in the same utterances by the English speakers. The remaining data involve partial loss of the coronal gesture, and are therefore susceptible to description within conventional phonological formalism. The weakening of coronal gestures in certain contexts appears only as an arbitrary stipulation within the theory of Articulatory Phonology. It is argued that the theory requires further elaboration to allow the behavior of the coronals to be modelled adequately.

1. CORONALS IN CC CLUSTERS

A number of studies have drawn attention to the tendency of alveolar and dental stops and nasals to assimilate to the place of articulation of a following non-coronal obstruent. The process is attested as source of phonological change in many languages, and gives rise, for example, to the presence only of homorganic intramorphemic NC clusters in English. The process has typically been formulated within the apparatus afforded by phonological theory in terms resembling those in figure 1, either, as in (a), in the linear formalism of early Generative treatments or as in (b), employing an autosegmental treatment of those features specifying place of articulation.

In this paper, however, I shall present evidence from rapid speech indicating that the formulations of fig. 1 are insufficiently revealing both of the phonetic facts obtaining in both English and Russian, and of the knowledge to which a native speaker of either language must have access in order correctly to produce sequences such as those under discussion.

2. ALVEOLARS IN ENGLISH

I have reported [1] an investigation into CC clusters in rapid speech in English, where C1 is an alveolar stop or nasal and C2 a velar stop, with an intervening mor-epheme or word boundary. Qualitative examination of electropalatographic (EPG) contact data for several speakers reveals a large number of utterances in which the coronal gesture is significantly reduced in magnitude, such that complete closure is not attained during the consonant. Speakers appear to differ in their choice of articulatory strategy here: the three options seemingly available are: (i) to execute a full coronal gesture, giving rise to full alveolar closure; (ii) to execute a weakened coronal gesture, with no complete closure; and (iii) to execute only the following velar gesture. While tokens of type (iii) are those which may be mod-

ded in conventional phonological descriptions as an assimilation, as in fig. 1, it is those of type (ii), exemplified in fig. 2, which, insofar as the forms they manifest are under the speaker's deliberate control rather than as the natural consequences of the inertial properties of the speech apparatus, must pose problems for conventional phonological rules and representations. This is because in these cases the coronal gesture involves a degree of lingual displacement, and perhaps also a duration, inconsistent with the discrete categories of binary feature-value and of timing-slot provided by theory.

3. QUANTITATIVE INVESTIGATIONS OF ARTICULATORY GESTURES

Further insight into patterns of articulatory activity may be gained by a consideration of the trajectories of individual articulatory subsystems, recently restored to the phonetician's armoury through the development of the concept of the gestures in the paradigm of Articulatory Phonology developed by Brownman and Goldstein [3]. In the work reported in the present paper, gestural trajectories were approximated from time-varying summaries of EPG contact data, and a number of measures devised by which temporal aspects of the various articulatory strategies might be compared. Figures 3 and 4 show gestural trajectories for the nasal + plus stop sequence [ng] in the phrase hand grenade. From the data values were obtained for (a) the duration of the alveolar and velar closures (DAC, DVC); (b) the overall duration of the coronal and dorsal gestures (DCG, DDO); (c) the degree of lingual displacement, corresponding to the height of the peaks for the two gestures (CMAX, DMAX); and (d) the interval between the peaks of the two closures, or, in the case where no velar closure was formed, between the peak in the coronal gesture and the onset of velar closure (INT).

In comparison with the slow utterance, for the fast utterance (fig. 4) CMAX is reduced to 70% of its maximum possible value, DCG is reduced by 10%, and DAC is zero; that is, the coronal gesture is diminished in magnitude to such an extent that no closure is formed, and also somewhat in duration. DMAX remains constant at 100%, DVC increases by 78% and DDO increases by 43%; the velar stop is fully articulated, and now significantly longer. INT is now ~11 ms: the velar closure is formed before the coronal gesture reaches its peak. Note also that the dorsal gesture is initiated before the coronal gesture. The data rise therefore a partial implementation of the restructuring implied by the autosegmental treatment of fig. 1b: the place of articulation originally associated only with the velar stop has 'spread' to occupy the conso-

Figure 2: EPG contact pattern for a weakened alveolar gesture.

Note that for the speakers investigated the final [d] in hand was usually elided in fast speech, and that the present investigation is confined to lingual gestures and hence has nothing to say about the tongue retraction and raising the assimilation associated with [s] is retained even when the coronal gesture is lost altogether, giving rise to a velar nasal [g].
DENTALS IN RUSSIAN

A consideration of the behaviour of speakers of Russian in similar contexts reveals some significant differences. The sound system of Russian differs from that of English in two significant respects: in general the requirement that NC clusters should be homorganic within the morpheme does not apply; and there is no surface contrast between dental and velar nasals. A large body of data from two speakers of Russian was subject to the same qualitative and quantitative investigation as the data from English. To begin again with qualitative observations, two points are immediately evident:

(i) in the case of CC clusters where C1 is a stop, no reduction can be observed in the magnitude of the coronal gesture as speaking rate increases (CMAX remains constant at 100%);
(ii) the range of contexts in which complete assimilation (i.e., a velar nasal) is encountered is very narrow, and apparently not sensitive to speech-rate. The cases involved were words such as /sanka/ and /funktsia/, in which the nasal and the following stop must be syllabified together (since the sequence /ks/ is impermissible as a syllable-onset). These forms showed [g] even in slow, careful speech.

In the remainder of cases, where the n and the following stop are heterosyllabic, the forms recorded typically reveal a fully articulated dental nasal in slow speaking and in fast speech a reduction in the magnitude of the coronal gesture, generally leading to the absence of a complete dental closure.

Applying the same quantitative measures as for English to the Russian data reveals further cross-linguistic differences. In the fast speech examples from the Russian speakers in the experiment, the reduction in magnitude of the coronal gesture is not accompanied by a corresponding lengthening in the duration of the dental gesture (CMAX decreases but DGD remains constant, or even undergoes a slight reduction typical at increased rates of speech), and while C1 decreases, the velar closure is nonetheless formed after the peak in the coronal gesture. Thus while the phonological formulation of fig. 1b was seen to be roughly appropriate to the articulatory patterns found in English, with weakened alveolars and lengthened velars suggesting a partial implementation of the phonological processes of autosegmental devoicing and spreading, no such interpretation appears suitable for the patterns found in Russian-speakers.

It is appropriate instead, I would argue, to view the weakening of the Russian dentals as the manifestation of a process more phonetic than phonological; that is, more representative of the natural constraints acting on the articulatory apparatus than of the principles of phonological organisation which may be discerned in the English found in Ohala's study. [4] If that is phonological pattern (a "sound change" in a diachronic perspective) has a phonetic motivation it is reasonable to expect to find evidence of the relevant phonetic process in speech production. Thus diachronic evidence of the instability of coronals in CC clusters leads us to expect a phonetic process of the sort encountered in the Russian data.

It would be inaccurate, however, to attribute the variety of weakened coronal gestures to the operation of a freely-applying natural phonetic effect: there is evidence that the phonetic form of coronals in English is not determined absolutely by the nature of the following stop, but that the degree of assimilation may vary depending upon various factors. The fact that the coronal stops in Russian are robustly resistant to weakening suggests at least that a particular phonetic effect may be blocked as part of the native speaker's low-level phonetic knowledge.

5. LEVELS OF PHONOLOGICAL KNOWLEDGE

We are therefore led to a picture of the organisation of the various types of knowledge of pronunciation, in which the variety of forms encountered in the data in this study are governed by principles operating on several levels:

- High-level phonological rules (cf. lexical rules)
  - Expressive in conventional phonological formalisms
    - e.g., distribution of Russian [g]: intra-phonemic NC clusters in English
- Low-level phonological rules (cf. postlexical rules)
  - Parallel implications not expressible e.g., English alveolar C1 in CC clusters across morpheme boundaries
- Phonetic effects
  - Phonetically motivated articulatory processes; may be phonologically blocked (e.g., Russian [t,si] or may apply freely (e.g., Russian [ni])

Two important consequences emerge: that some aspects of the speaker's knowledge of how their language is pronounced involve 'mixed' forms which conventional phonological theories are not equipped to represent; and that language-specific knowledge of pronunciation extends to the operation or blocking of natural low-level processes.

6. CORONALS IN ARTICULATORY PHONOLOGY

The paradigm of Articulatory Phonology [3] appears well-equipped to accommodate a variety of low-level phonetic detail which, as I have argued, falls within the subject-matter of a comprehensive theory of phonology. Gestural organisation is linked to high-level phonological representations, and the operation of the task-dynamic model yields a spatio-temporal representation in terms of gestural trajectories in which the non-discrete application of phonetic and phonological processes may be formalised. In addition, the application of gestural principles governing relationships of phase between gestures accounts for much of the data we have observed, in which the velar gesture is responsible for the 'masking' of the coronal gesture.

What is still lacking in current formulations of the theory is a convincing account for the facts of coronal-gesture weakening. That gestures weaken in casual speech is stipulated somewhat axiomatically, and in no sense can be said to emerge from the mathematical properties of the model. Moreover, there appears to be no way, in a model which treats all gestures as formally identical objects, in which it can be shown that coronal gestures specifically are subject to elision in CC clusters. At the heart of the matter is the modelling of gestures as the critically-damped attraction of the active articulator towards its target. Thus for an articulator to fall during its target during the execution of a gesture seemingly requires the target itself to be reprogrammed. Within existing versions of the theory it would seem to be necessary to abandon the assumption of critical damping (such that an articulator always reaches its target) in order to accommodate gestural weakenings, and other undershoot phenomena. A more drastic revision of the model would be to abandon the modelling of gestures in terms of attraction, in favour of a model in which the articulator is pushed rather than pulled towards its target. But this would be to abandon entirely the mathematical content of the theory.

The issue of gestural weakening clearly remains a problem for the development of the theory; it seems clear that evidence of the kind presented in this paper will be of relevance in seeking a solution.

REFERENCES

ARTICULATION OF PROSODIC CONTRASTS IN FRENCH

J. Fletcher * and E. Vaikliotis-Bateson **

* Speech, Hearing and Language Research Centre, Macquarie University, Sydney, Australia, and ** ATR Visual and Perception Research Laboratories, Kyoto, Japan

ABSTRACT

The current study examines the influences of intonation and syllable structure on accentuation and final lengthening in a corpus of articulatory data. While consistent kinematic patterning across speakers was not observed for intonation differences, it is apparent that different articulatory manoeuvres are employed to bring about accent-related duration change in open and closed syllables.

1. INTRODUCTION

Many studies of the acoustic correlates of accentuation in French have examined this phenomenon in syllables at the edge of major prosodic phrases or sentences (e.g. Delattre [1]; O'Shaughnessy [2]). More recent investigations (e.g. Touati [3]) separate the two classes of accented syllable (accented final and accented as final), and note that accent-related duration differences are somewhat reduced in the phrase-final context.

In a recent paper (Fletcher and Bateson [4]), we propose that accentuation and phrase-final lengthening are associated with different underlying articulatory manoeuvres. As suggested by Edwards et al. [5] for English, final lengthening in French involves a specific lengthening at the phrase-edge. Accentuation, by contrast, is a change in linguistic prominence and not essentially a duration contrast. The two linguistic phenomena should not be confused in experimental designs.

In the current study, we re-examine the phrase-internal accented/unaccented contrast in a corpus of articulatory data, based on natural as opposed to reiterant speech. An extra "level" of accent is also examined by comparing pretonic accented syllables with tonic accented syllables (syllables associated with a melodic peak). We also look at the influence of tone and syllable structure on the articulatory timing of phrase-final syllables. In an early acoustic timing study of accent in French, Benguerel [6] claims that accentual lengthening is greater when intonation is falling rather than rising. He also claims that the lengthening effect is strongest in open as opposed to closed syllables. It is of interest to see how these effects manifest themselves in the underlying articulation of syllables.

2. METHOD

Two speakers of French produced ten repetitions of the sentences shown in Table 1 at two self-selected tempi, conversational normal and fast. The sentences were devised in such a way that the test tokens (indicated in uppercase) represent different prosodic categories. Set A places the tokens (chosen to contrast open and closed syllables) in unaccented (PAPA) pretonic, accent (PAPE), and tonic accent contexts. Set B places the tokens in sentence-final declarative and sentence-final interrogative contexts. In all instances, the token in the sentence B (i) was recited with a low, slightly falling tone.

<table>
<thead>
<tr>
<th>Set A</th>
<th>Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>Le PAPA: a par lui</td>
</tr>
<tr>
<td>(ii)</td>
<td>Le PAPA: a par lui</td>
</tr>
<tr>
<td>(iii)</td>
<td>Le PAPA: a par lui</td>
</tr>
</tbody>
</table>

The token in sentence B (ii) was recited with a rising tone, commonly associated with a yes/no question.

Vertical movements of the lower lip, upper lip and jaw were recorded using the modified SELSPOT opto-electronic articulator tracking device at Haskins Laboratories. The digitized and low-pass-filtered position signals were corrected for any head movement and were numerically differentiated to produce instantaneous velocity. Vertical position of the lower lip was subtracted from that of the upper lip to obtain lip aperture. Peaks in the movement trace (Fig. 1) correspond to points of maximum closure associated with the production of the bilabial consonant and valleys correspond to maximum opening associated with the production of the low back vowel.

![Figure 1: Kinematic Measures](image)

Measurements of gesture duration, displacement, and associated peak velocity using automatic peak picking were noted for opening gestures in the case of /pa/ syllables, and for both opening and closing gestures for /pa/ syllables. The time course of gesture velocity was also examined. We are calling the time period from the onset of the gesture (defined as the last point of zero velocity before the opening or closing gesture) to the time when peak velocity is registered in the gesture, the acceleration phase, and the time period from the peak moment to the offset of the gesture, the deceleration phase, in accordance with earlier work by Nelson [7] among others.

3. RESULTS

The results of the kinematic analysis are presented in Tables II and III. All results of within group comparisons (Kirk [8]) cited in the following paragraphs, are significant at p<0.01. For subject AS, tonic accented /pa/ syllables have significantly longer opening gestures and bigger lip apertures than unaccented /pa/ syllables (F's 6.15, 15.19), with no significant differences in peak velocity. By contrast, acceleration and deceleration duration and longer in the opening gestures of accented compared to unaccented syllables (F's 48.25, 11.13). By contrast, speaker BA, shows no overall duration contrast, but unaccented /pa/ opening gestures are significantly bigger and faster than tonic accented gestures (F's 8.78, 28.69).

For the pretonic/pretonic contrast in /pa/ syllables, there are no significant duration differences in opening gestures for either speaker. Conversely, closing gestures in tonic accented syllables are consistently longer than pretonic gestures (AS:F, 10.59; BA:F, 9.44). This difference is localised to the deceleration portion of gestures for both speakers (AS:F 6.58; BA:F, 5.44). Tonic syllables also have bigger opening
and closing lip apertures in BA’s data (F’s 14,44,16,79) coupled with higher peak velocities (F’s 13,71, 7,96). AS shows no significant lip aperture differences, but peak velocities are lower in closing gestures of tonic syllables (F’s 3,63).

**TABLE II** - Mean and standard deviation values (in parentheses) of opening and closing velocity differences (ms/s) and peak apertures (mm), peak velocity (mmms), acceleration and deceleration durations (ms) in /pap/ syllables (token - TAP).

<table>
<thead>
<tr>
<th>Unsecutive</th>
<th>Tonic</th>
<th>Final(Low)</th>
<th>Final(NS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. AS 101(8)</td>
<td>191(11)</td>
<td>169(8)</td>
<td>182(14)</td>
</tr>
<tr>
<td>BA 73(5)</td>
<td>72(9)</td>
<td>131(11)</td>
<td>138(12)</td>
</tr>
<tr>
<td>LA.</td>
<td>76(2,9)</td>
<td>9(1,2)</td>
<td>9(1,6)</td>
</tr>
<tr>
<td>Vp.</td>
<td>5,41(88)</td>
<td>8,31(83)</td>
<td>9,98(79)</td>
</tr>
<tr>
<td>Acc.</td>
<td>146(13)</td>
<td>152(16)</td>
<td>122(25)</td>
</tr>
<tr>
<td>BA 169(30)</td>
<td>134(24)</td>
<td>119(21)</td>
<td>153(35)</td>
</tr>
<tr>
<td>Acc.</td>
<td>63(3)</td>
<td>103(8)</td>
<td>84(11)</td>
</tr>
<tr>
<td>BA 43,08</td>
<td>42(4)</td>
<td>75(6)</td>
<td>70(7)</td>
</tr>
<tr>
<td>Decoel</td>
<td>35(6)</td>
<td>67(10)</td>
<td>84(15)</td>
</tr>
<tr>
<td>BA 20(2)</td>
<td>50(4)</td>
<td>56(6)</td>
<td>64(10)</td>
</tr>
</tbody>
</table>

**TABLE III** - Mean and standard deviation values (in parentheses) of opening and closing duration differences (ms), lip aperture (mm), peak velocity (mmms), acceleration and deceleration durations (ms) in /pap/ syllables (token - TAP).

<table>
<thead>
<tr>
<th>Unsecutive</th>
<th>Tonic</th>
<th>Final(Low)</th>
<th>Final(NS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. AS 121(8)</td>
<td>136(13)</td>
<td>127(9)</td>
<td>107(4)</td>
</tr>
<tr>
<td>BA 69(4)</td>
<td>76(5)</td>
<td>93(4)</td>
<td>114(9)</td>
</tr>
<tr>
<td>LA.</td>
<td>94(5)</td>
<td>71(3)</td>
<td>91(7)</td>
</tr>
<tr>
<td>Vp.</td>
<td>7,8(9)</td>
<td>10,2(5)</td>
<td>10,1(2)</td>
</tr>
<tr>
<td>Acc.</td>
<td>152(22)</td>
<td>146(13)</td>
<td>161(9)</td>
</tr>
<tr>
<td>BA 173(28)</td>
<td>151(12)</td>
<td>173(16)</td>
<td>160(17)</td>
</tr>
<tr>
<td>Accel.</td>
<td>62(9)</td>
<td>69(14)</td>
<td>69(8)</td>
</tr>
<tr>
<td>BA 49(5)</td>
<td>49(5)</td>
<td>60(7)</td>
<td>68(8)</td>
</tr>
<tr>
<td>Decel.</td>
<td>59(6)</td>
<td>67(5)</td>
<td>58(7)</td>
</tr>
<tr>
<td>BA 30(1,5)</td>
<td>33(4)</td>
<td>32(3)</td>
<td>46(1,2)</td>
</tr>
</tbody>
</table>

Only speaker BA shows significant kinematic differences according to tone. In /pap/ syllables, opening and closing gestures are longer when tone is rising (F’s 59,36, 17,37) than when tone is low. This duration difference is reflected in both the acceleration and deceleration portions of opening gestures of /pap/ syllables (F’s 6,23, 35,99) and the acceleration portion of /pap/ closing gestures (F,47,99). There are no tone-related lip aperture differences or significant peak velocity differences in /pap/ opening gestures, although closing gestures are slower when tone is rising (F,3,99). No significant duration differences are observed in /pap/ gestures although lip aperture is bigger and peak velocities higher in syllables with rising tone (F’s 8,8, 4,17).

**4. DISCUSSION AND SUMMARY**

Clearly, more data are needed to supplement this initial analysis, especially in view of the degree of inter-speaker variability. Some generalisations can be made, however. As in our earlier study, these data suggest that more than one type of articulatory manoeuvre underlies these prosodic contrasts. Conventional accent or stress effects - longer, bigger gestures - are evident in /pap/ syllables for speaker AS, and /pap/ syllables for BA. It can also be argued that the observed bigger apertures in word initial /pap/ syllables for AS are also an accent effect, given the increased predominance of word initial accent in spoken French. Speaker BA consistently accent the first syllable of "Papa" in sentences (i) and (ii). The localisation of the duration contrast to the tailend of closing /pap/ gestures suggests that protracted closing gestures may be cut short by the opening gesture associated with the upcoming syllable in the sequence. In other words, gestural slurring, resulting in truncation of closing gestures may explain shorter duration differences in pretonic syllables (Saltzman and Munhall[9]). In addition, changes in underlying amplitude of both opening and closing gestures may determine observed kinematic patterning in BA’s tonic accented productions and AS’/pap/ data.

In AS’/pap/ data, on the other hand, the lack of a lip aperture difference, coupled with slower peak velocities suggest alteration of another underlying control variable - i.e. gesture stiffness, or force (Saltzman and Munhall[9], Edwards et al.[5]) without a change in underlying gesture amplitude. This latter pattern does not suggest a typical stress or prominence contrast for this syllable. It is more like the pattern for final lengthening noted by Edwards et al. for English.

While results for the tone contrast are not consistent across speakers, they suggest that syllables associated with rising tone are as long or longer than syllables associated with falling tone, contrary to Benguerel’s claims. Duration effects are clearest in closed as opposed to open syllables. The lack of lip aperture differences and slower peak velocities in rising tone /pap/ syllables again suggest a similar articulatory manoeuvre to that noted for final lengthening in English by Edwards et al. By contrast, the bigger lip apertures and higher velocities in rising tone /pap/ syllables without an accompanying duration difference suggest an articulatory manoeuvre not unlike that attributed to a stress contrast.

**5. REFERENCES**


Acknowledgments:

Parts of this research were supported by the National Science Foundation (USA) under grants no. IRI-8858109, IRI-861785 to Mary Beckman, the Ohio State University and by the National Institutes of Health (USA) under grant no. NS-13617 to Haskins Laboratories.
ESSAI DE MÉTHODE POUR LA RECHERCHE DE L'IMAGE CENTRALE : VOYELLES [i, e, a] DU FRANÇAIS.

Annie Pastor
Institut de Phonétique - Université de Strasbourg II
22 rue Descartes - 67084 Strasbourg Cedex - France

ABSTRACT

In this contribution, we study the articulatory realization of three French vowels [i, e, a] placed at the end and in the middle of rhytmical groups. We use X-Ray films for one speaker and we choose 14 parameters. The results show that it is more difficult to find a central image when a parameter is stable. Our second intention is to establish a hierarchization of parameters based on the part they play to help finding the central image.

1. BUT ET MÉTHODE

Notre étude porte sur l’exploitation de films radiologiques avec synchronisation image/son (50 images par seconde) 1112. Nous retenons 14 paramètres (fig. 1) :

1 et 2 : projection de la lèvre supérieure et inférieure.
3 : écartement labial.
4 : angle des maxillaires. 5.6.7.8.9 : hauteur de la langue.
10 : racine de la langue.
11 : hauteur maximale du voile du palais.
12 : os hyoïde (mouvement vertical et horizontal).
13 : base du larynx.
14 : épiglotte (mouvement horizontal).

Nous relevons le début et la fin acoustique de chaque voyelle ainsi que la dernière image de la consonne qui la précède et la première de celle qui la suit (position interconsonantique).

Dans notre corpus, nous relevons en fin de groupe rythmique : 17 [i] , 6 [e], 14 [a], en milieu de groupe rythmique : 7 [i], 1 [e], 10 [a]. Les voyelles sont précédées de consonnes suivantes : [p, b, f, s, k, g]. Il nous faudra tenir compte du contexte qui suit. Dans un premier temps nous sommes intéressés à l’exploitation de paramètres correspondant à une période de stabilité qui les caractérise [3] [4]. Nous avons relevé les mesures de la durée totale de la voyelle. L’analyse du comportement détermine les paramètres qui servent d’indices pour dégager l’image centrale.

2. ANALYSE

Illustrons ceci par un exemple : Phrase 19 (sibota’pi) Il s’agit du [i] en position interconsonantique (fig. 2). Nous choisissons l’image qui suit le moins l’influence du contexte, progressive et régressive. Etablissons chaque paramètre :

Les lèvres : dans les deux cas, nous relevons une période de stabilité de trois images (14 16). Nous savons que pour [s] les lèvres demeurent étirées comme pour [i]. En revanche, sous l’influence de la syllable suivante [bo], la projection labiale s’intensifie. Par 3 : courte période de stabilité où l’écartement labial est maximal à 11.5 mm (images 14 et 15). Sous l’influence de la consonne bilabiale suivante, les lèvres vont très vite se refermer.

Par 4 : nous notons une très faible variation de l’angle des maxillaires. Le mouvement général correspond à une ouverture de l’angle en raison du contexte qui suit. Nous retenons la période de stabilité qui se situe au centre de la durée de la voyelle à 1.5 mm (images 15 et 16).

Par 5.6.7 : Ces paramètres correspondent à la partie antérieure et centrale de la voyelle. Nous constatons que celle-ci s’élève dans les trois cas et que nous retenons la période de stabilité où la hauteur de langue est maximale : Par 5 :45 mm (images 15 et 16) ; Par 6 et 7 : 48 mm (images 15 à 17).

<table>
<thead>
<tr>
<th>Images</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Param.</td>
<td>67</td>
<td>67</td>
<td>67</td>
<td>68</td>
<td>69</td>
</tr>
<tr>
<td>1</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td>67</td>
<td>68</td>
</tr>
<tr>
<td>2</td>
<td>11.5</td>
<td>11.5</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.5</td>
<td>1.5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>43</td>
<td>45</td>
<td>45</td>
<td>43</td>
<td>39</td>
</tr>
<tr>
<td>5</td>
<td>48</td>
<td>48</td>
<td>45</td>
<td>48</td>
<td>45</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>47</td>
</tr>
<tr>
<td>7</td>
<td>43</td>
<td>43</td>
<td>44</td>
<td>43</td>
<td>46</td>
</tr>
<tr>
<td>8</td>
<td>35</td>
<td>35</td>
<td>35</td>
<td>38</td>
<td>41</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>28</td>
<td>28</td>
<td>30</td>
<td>31</td>
</tr>
<tr>
<td>10</td>
<td>68</td>
<td>68</td>
<td>68</td>
<td>68</td>
<td>68</td>
</tr>
<tr>
<td>11</td>
<td>(13;36')</td>
<td>(13;34')</td>
<td>(13;34')</td>
<td>(13;34')</td>
<td>(13;33')</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>35</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>34</td>
</tr>
<tr>
<td>14</td>
<td>35</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fig. 1</th>
</tr>
</thead>
</table>
Nous choisissons les mesures qui rendent compte de ce comportement à 35 mm (images 14 et 15). Puis nous relevons que la racine de la langue se rapproche progressivement de la paroi pharygale sous l'influence du contraste de la voyelle vélaire [o].

Par 11 : Le voile du palais demeure quant à lui parfaitement stable pendant la durée totale de la voyelle à 68 mm.

Par 12 : Dans cet exemple, l'os hyoïde est uniquement mobile sur le plan vertical. Nous choisissons le moment où il se stabilise sur ce plan. Cette période correspond aux deux images 14 et 15 à 36 mm.

Par 13 : Les mesures de la base du larynx ne varient que d'1 mm. Notre choix se porte sur la période de stabilité centrale à 8 mm (images 15 à 17).

Par 14 : L'épiglotte suit le mouvement de la racine de la langue. De ce fait nous sélectionnons les images dont les mesures correspondent au moment où elle se situe le plus loin de la paroi pharygale à 33 mm (images 15 à 17).

3. DISCUSSION

3.1 Paramètres - indices

Une image se dégage nettement : l'image 15. Elle apparaît comme le point commun de toutes les périodes de stabilité relevées. Par ailleurs, c'est à cette image que la voyelle subit le moins les influences voisines. Il s'agit d'un mouvement de la consonne bilabiale [b] en ce qui concerne l'écartement et la projection des lèvres, ainsi que la voyelle vélaire [o] pour la langue (principalement la partie postérieure), l'angle des maxillaires et l'os hyoïde qui s'élève.

Parallèlement certains paramètres nous ont aidée à déterminer l'image centrale. Ils se caractérisent par une période de stabilité courte : Par. 1, 2, 3, 4, 5, 8, 9, 10, 12, 13, 14. Les autres, peu nombreux pour cet exemple, ne nous offrent pas d'information particulière en raison de leur trop grande stabilité : Par 6, 7, 11. Nous ne pouvons établir de hiérarchisation type en ce qui concerne les voyelles en milieu de groupe rythmique par la trop grande influence du contexte. En revanche, en fin de groupe rythmique nous pouvons en établir une. Le classement se présente comme suit :

- os hyoïde : période de stabilité très courte pour le mouvement à la fois horizontale et verticale.
- partie antérieure et centrale de la langue : Par 5, 6, 7 (hauteur maximale).
- racine de la langue : rapprochement où éloignement maximal.

Quant aux autres paramètres ils ne détiennent pas autant d'information de par leur grande stabilité (partie postérieure de la langue : Par 8, 9 ; voile du palais ; épiglotte) ou mobilité : base du larynx.

3.2 Place de l'image centrale

3.2.1 Voyelles en milieu de groupe rythmique

En nous référant à l'exemple ci-dessus, nous constatons que l'image centrale se situe en début de voyelle. Cet exemple constitue une exception comparativement aux autres exemples étudiés. En effet, l'image centrale correspond au milieu de la durée des voyelles [i, e, a] confondues. La durée varie de 10cs à 16cs pour [i] et [a] et de 12cs pour [e]. En ce qui concerne durée et place de l'image centrale, nous ne retenons pas de différence notoire entre [i] et [a].

3.2.2 Voyelles en fin de groupe rythmique


Comme nous le constatons dans ce tableau (fig.3), l'image centrale se situe après le milieu de la durée de la voyelle. Il est évident que plus la voyelle s'allonge plus l'image centrale se décale vers la fin de la voyelle.

Enfin, la comparaison entre voyelles en fin et en milieu de groupe rythmique met en évidence une diminution de 33,33 % pour [i] par rapport à [i]; de 40 % de [e] par rapport à [e]; de 45 % de [a] par rapport à [a].

4. CONCLUSION

L'étude des voyelles [i, e, a] nous a permis de montrer que l'image centrale se situe au centre du milieu de la durée de la voyelle en position interconsonantique et après pour les voyelles en fin de groupe rythmique.

Une hiérarchisation des paramètres-indices est uniquement possible pour les voyelles en fin de groupe rythmique. Courte stabilité et mobilité constituent les deux critères essentiels qui mettent en évidence l'image centrale. Nous avons souligné l'importance de la partie antérieure de la langue et de la racine, mais surtout celle de l'os hyoïde. Celui-ci ne joue pourtant pas de rôle primordial dans la réalisation articulatoire des voyelles.

Enfin, notre essai de méthode nous permet de connaître le moment précis où le contexte exerce son influence. L'analyse séparée des paramètres nous indique s'ils réagissent de manière identique ou différente ; avec rapidité ou retard. La variabilité intrinsèque de chacun d'eux ne pourra que confirmer les tendances.

5. REFERENCES


DE L'ANALYSE D'UNE VARIATION DE DÉBIT DANS LA CHAINE PARLÉE, A LA LUMIÈRE DE LA CINERADIOGRAPHIE

Béatrice Vaxelaire

Institut de Phonétique - Université de Strasbourg II
22 rue Descartes - 67084 Strasbourg Cedex - France

ABSTRACT

The aim of this work is to evaluate in the light of cineradiography, the articulatory behavior of stop consonants used in fast speech. This paper describes particularly the unvoiced stop consonants [p, t, k], unstressed, at the intervocalic position with identical environment, with the single and successive double consonant at two different rates. It results through these first measures of our study, fast rate implies a few reduction of the articulatory gestures and some compensatory interarticulator gestures.

1 BUT ET METHODE

1.1 Présentation


1.2 Corpus.

Phrases retenues et segments étudiés :
1. Il a pas mal. [ilapa'mal]
2. Les atabler. [lezata'bler]
3. Très acarâtre. [tazaka'atra]
4. Il zappe pas mal. [ilazappa'mal]
5. La chatte tachetée. [la'katta'te]
6. Trois sacs carrés. [tw asakka're]

1.3 Paramètres.

Nous avons relevé 14 paramètres (fig. 1) et 2 : projection des lèvres supérieure et inférieure.
3 : écartement labial.
4 : angle des maxillaires.
5,6,7,8,9 et 10 : langue.
11 : voile du palais (hauteur maximale, hauteur et écartement de la paroi pharyngale du creux dans la partie postérieure et inférieure du voile, distance d'occlusion avec la paroi pharyngale).
12 : os hyoïde (mouvement horizontal et vertical).
2.2 Comportement temporel.
Les phasmes du corpus (1.2) voient leur durée réduite entre le lento et l’allegro en moyenne de 30%. La durée des consonnes étudiées réduit en moyenne de 27% sauf pour l’extrabucale double successive (ph. 4) où la réduction temporelle est de 52,5%. En général la durée des consonnes simples réduit moins que celle des doubles successives (respectivement 23% et 33%). C’est pour [p] aussi bien en simple qu’en double que la durée diminue le plus. [t] et [k] ont une réduction identique en simple, mais elle est plus importante pour [k] en double.

3. ANALYSE DES MESURES.
La perturbation de débit provoque un certain nombre de modifications articulatoires : 
Par. 1 et 2 : les lèvres sont moins proéminentes en allegro avec une position générale plus arrière de 3 mm pour la consonne double successive en allegro.
Par. 3 : la durée d’occlusion de l’extrabucale est réduite en allegro (de 20cs à 10cs pour [pp] l’écart de d’occlusion est supérieure en allegro (de 2mm). L’écartent bilabial est plus marqué en allegro.
Par. 4 : L’écartement du maxillaire est inférieur en allegro (de 3mm, [pp] de 4mm, [k] et [kk] de 2mm).
Par. 5 : La distance d’occlusion pour l’alvéolé dentaire est supérieure en allegro (de 2mm).
Par. 6 : partie de la langue plus élevée en allegro (2mm).
Par. 7 : partie de la langue plus élevée en allegro, surtout pour la velaire (2mm) et l’alvéolé dentaire (4mm).
Par. 8 : l’occlusion est retrouvée de 2cs en allegro pour la velaire. Pour [kk] la durée d’occlusion est supérieure de 3cs en allegro.
Par. 9 : partie de la langue plus élevée en allegro (2 à 3mm).
Pour les paramètres 6 à 9, la langue est plus élevée pour la consonne double, en moyenne de 5mm dans les 2 débits.
Par. 10 : le mouvement de la racine est réduit en allegro et décalé vers la paroi pharyngale pour [t] et [tt] de 2 à 5mm. Les mesures sont décalées vers l’avant en allegro de 1mm.
Par. 11 : le sommet du voile est plus élevé en allegro (sauf [t]) de 3mm pour [tt, pp] la distance d’occlusion est supérieure en allegro (sauf [pp]) de 4mm pour [t, t], 7mm pour [tt], 6mm pour [k] et 3mm pour [kk]. Nous observons un creux dans la partie inférieure postérieure du voile pour [s] et [k]. La hauteur du creux est supérieure en allegro de 3 et 4mm respectivement pour [p] et [k]. L’écartement pharyngal est supérieur en lento de 7mm pour [k]. Nous n’observons pas de creux pour [t] ; cependant nous remarquons un rapprochement du voile avec la paroi de 2mm en allegro.
Par. 12 : l’os hyoïde est plus reculé en allegro pour les consonnes doubles successives de 2 à 3mm. Il est en moyenne plus bas en allegro de 2mm (sauf [tt]).
Par. 13 : la base du larynx subit un abaissement en moyenne de 2mm, sauf pour les consonnes doubles en lento de 4mm. Ce mouvement est décalé vers le bas pour [t] de 5mm en lento et 2mm en allegro.
Par. 14 : l’épiglotte reculée vers la paroi en moyenne de 6mm (3mm pour [k]) et se rapproche de la paroi pour [kk]. Nous observons un décalage des mesures en allegro vers la paroi pour [t] et vers la racine pour [k].
Commentaire : le débit est abordé ici comme variable articulatorie. Des différences de comportement des articulateurs peuvent être liées à des paramètres articulatoires. Nous observons un décalage des mesures en allegro vers la paroi pour [t] et vers la racine pour [k].

4. CONCLUSION.
La chaîne articulaire est un système de perturbation de débit montre au-delà d’une réduction de la durée, une modification du schéma dynamique des articulateurs. Notre recherche nous permet de déterminer certaines résistances des articulateurs à la variation de débit, et de mettre en évidence une stratégie des articulateurs en rapport avec le débit. Nous avons observé un phénomène de réduction articulatorio touchant plus ou moins certains articulateurs, et provoquant un phénomène de compensation interarticulatoires.

REFERENCES.

Nos sincères remerciements à Pierre-Yves Connan qui nous a aidé à réaliser la mise en page de cet article.
PHONOLOGICAL ORGANIZATION IN BILINGUALS:
EVIDENCE FROM SPEECH ERROR DATA

CHENEY CROW

Dept. of French & Italian
University of Texas, Austin, TX 78712, USA

ABSTRACT.
Effects of bilingualism on phonological organization were investigated by comparative analysis of speech errors in late French/English bilinguals, 10 native speakers of each language. In comparison with (10) monolingual controls in French and English, some error categories were consistent with existing data, while significant differences in other categories previously considered "universal" were observed in all bilinguals.

1. INTRODUCTION
One aspect of bilingual speech which has not been investigated is the phonological organization of speech production. Speech errors are considered evidence of at this level of phonological organization; speech error behavior has been taken into consideration in most current models of speech production (Fowler, 1987). Nearly a century of analysis of spontaneous, and more recently, elicited, speech errors in German, English, and Dutch have revealed regularities in certain characteristics of speech errors (reviewed in: Cutler, 1982). Speech errors of aphasics have also demonstrated the same, consistent pattern (Blumenthal, 1990).

Speech error behavior in bilinguals has not been investigated. As significant differences between the first and second languages of late bilinguals have been observed in many aspects of speech behavior, it was hypothesized that speech error analysis could reveal differences in the phonological organization of speech production between the first and second languages of late bilinguals. The prediction was that speech errors of bilinguals would not indicate independent behavior of segments unique to the second language, and that no error would violate phonotactic constraints of the first language.

Initial results indicated significant differences between phonological organization in both languages and monolingual speakers of their first languages, as well as differences between the two monolingual groups. These differences were mainly examined, for they included "violations" of characterizations previously considered universal in speech error behavior.

2. PROCEDURE
A speech-error elicitation task, modeled on one created by Sheehy-Huffner (1987), was designed to elicit speech errors from monolingual and bilingual speakers of French and English.

2.1 Subjects
Four subject groups were chosen: (1) 10 monolingual French speakers; (2) 10 monolingual English speakers; (3) 10 native speakers of French, late bilinguals in English; (4) 10 native speakers of English, late bilinguals in French. Late bilinguals were chosen because of the evidence of significant differences observed between early and late bilinguals in second language competence, performance, and cortical behavior (Vaid 1987). All bilingual subjects had lived in a country in which the second language was spoken for periods of more than one year, and at the time of testing used both languages daily. All rated themselves as fluent speakers of their second languages.

2.2 Method
Four word sets comprised of two monosyllabic and two disyllabic words were presented to subjects in each language. All words were consonant initial, and varied in syllable structure from CVC to CVVCV structure. 35 of the word sets had sound sequences which were possible in both languages, with segments which exist in both languages. Syllable structure was the same in the two sets. Examples: English: fame fade pool; French: parade fade foole. The remaining five word sets were different in the two languages. These did not include the same syllable structure. All sets included segments unique to each language and in word-onset position. Example: (Target segment: TH) English: six thick jibbates sticks.

Subjects were presented with index cards on which the four-word sets were printed. Subjects were instructed to read each card three times, then to set the card down and repeat the four-word set from memory three more times, for a total of six repetitions. To avoid a memory confound, subjects were instructed to refer to the card if necessary during the final three repetitions.

Monolingual subjects were recorded in a single session. Bilingual subjects were recorded in separate sessions for their two languages, at a minimum interval of three weeks, because of the similarity of the two stimulus sets.

2.3 Data Analysis.
All sessions were transcribed, and errors were classified in several ways. Counts were made of consonant, vowel, word order and blend errors. These were further classified as either exchange, replacement, intrusion, or deletion errors. Position in word for all errors was recorded.

For interaction errors, the substitutions and exchanges, in which both the target segment and the uttering segment involved in an error occurred in the word string, the direction of the error (either anticipatory or perseveratory) and the relative position in word of the target and the uttering segment in the speech error were recorded. Stress was also noted, for both the target and the uttering segments, as well as voicing and place of articulation.

For intrusions, in which the uttering segment in an error does not occur in the stimulus set, comparison was made between the target segment and the uttering segment for syllable structure, placement of articulation, and rhyme. The number of segments involved was recorded, and errors were examined for word formation. All errors, both interactions and intrusions, which resulted in word formation were compared to target words for rhyme and syllable structure.

Data analysis included counts of all error types for each subject. For all groups, total counts, calculations of means and standard deviations were made for all error types. Between-group comparisons were tested by ANOVA and Chi Square analysis.

3. RESULTS
Four main trends were observed:
1. Similarities between groups.
2. Significant differences between French and English monolinguals.
3. Effect of second language acquisition on error type, size and position, on both first and second languages of bilinguals.
4. Language-specific differences in segment repertoire.
3.1. Similarities between groups.

Several types of speech error categories were similar in all groups, and consistent with existing data. For these error types, significant differences were not observed either between or within subject groups. The categories for which this occurred were: (1) the ratio of anticipatory to perseveratory errors; (2) position effect — the ratio of interaction of segments sharing word-final position, to those in different word position (initial/initial to initial/medial, etc.); (3) stress effect — the ratio of interacting segment-bearing similar lexical stress to those bearing different lexical stress; and (4) the percentages of total errors for each group that were: anticipatory, perseveratory, exchange, replacement, and word order errors.

3.2. Significant differences in error rates for French and English monolinguals.

Unlike monolingual English speakers, who have demonstrated a clear bias towards word-initial position errors, monolingual French speakers made a large percentage of their errors (up to 70%) in word-final position. The rules affect consonants in word-final position in French: (1) final consonant deletion; (2) for coronals only: variability in production — word-final coronals are produced only if adjacent word is vowel-initial. These phonological properties of word-final consonants in French may influence this effect, as word-final errors in monolingual French speakers occur almost exclusively on coronals.

3.3. Effect of second language acquisition on error position, size, and type.

Bilingual native speakers of English produced up to 30% of their errors in both French and English, in word-final position. These errors were not dominated by coronals in word-final position. Like the errors of bilingual English speakers, word-final errors of French bilinguals were neither restricted to, nor dominated by, coronal consonants, in either French or French. These results indicate either interactive effects between the first and second languages, or an effect of bilingualism which creates an unrestricted bias toward word-final errors.

Error unit.

While errors of monolingual speakers involved units which varied from 1 to 5 segments, almost all errors by bilinguals involved segments only. The only errors of bilinguals which involved units greater than a single segment were "blend" errors, a type of errors where numbers from two words in the stimulus set, in the first language.

Error type.

a. Blends. Although "blend" errors were made by almost all monolingual speakers, very few blends were made by bilinguals, and all in their first language. No "blend" errors occurred in the second language of bilinguals. All L2 errors were restricted to a single segment, i.e., the blend occurred in the first language and was not extended to the second language.

b. Deletion. No deletions were made by monolingual speakers. Deletion errors were made only by bilinguals, only in French, and only on word-final consonants.

c. Intrusion.

Size. Intrusion errors made by bilinguals ranged from 1-5 segments in size, while intrusion errors were restricted to single segments.

Word formation. 93% of monolingual intrusion errors resulted in word formation. Words were formed by bilingual intrusion errors only in L1 (the native language). Rhymes. 82.5% of English monolingual and 90% of French monolingual intrusion errors created rhymes with target words. Bilingual intrusion errors did not create words which rhymed with the targets.

3.4. Language-Specific differences in Segment Repertoire.

The word-final errors of any type were made by any bilingual speaker in which a segment which was unique to the second language occurred as a substitution for any other target.

4. DISCUSSION.

The fact that some categories of errors occurred with similar frequency in all groups, corresponding to existing data on speech error behavior, may indicate that these aspects of speech error behavior are more "language-universal" than other categories. The differences, however, which might indicate that "universals" must be tested in more language populations, and speaker types (bilingual and monolingual) before they can truly be classified as invariable.

Monolinguals.

The difference in dominant error position between French and English monolinguals is interpreted as consistent with existing data. Because of the restriction of word-final errors to coronal consonants, these errors may be considered word-initial, as word-final coronals, when produced, resyllabify as onset consonants of adjacent vowel-initial words.

Bilinguals.

The differences in speech error behavior between bilingual and monolingual speakers indicate that second language acquisition affects the phonological organization of speech planning in both their first and second languages. The elements affected are: error position, size, and type. The characteristics of the word-final errors of both bilingual groups could be explained by interaction of the two phonologies. The other changes, error size and type, are more difficult to explain, and demand further investigation. Since the "modality" of a segment, its occurrence as a substitution for another segment in positions or words other than its target position, is considered in "independent" behavior, it might be concluded that L2-only segments did not function independently. Consequently, no need to process these segments may bring about a more "holistic" processing of second language words in which they occur. There is abundant evidence of right hemisphere participation in the processing of second language speech of bilinguals, which may involve a more "holistic" function (reviewed in Fabro et al., 1990). Further study of other bilingual populations is indicated to further explore the "universality" issue, and the effects of bilingualism.

REFERENCES.


COORDINATION DU GESTE ET DE LA PAROLE
DANS LA PRODUCTION D'UN INSTRUMENT TRADITIONNEL

V. Berthier, C. Abry, T. Lallouache
Institut de la Communication Parlé,URA CNRS n°368
Grenoble, France

ABSTRACT
This paper describes an early learned coordination between gesture and speech; during traditional whistle making, children could utter rhymes. In the present case study, it appeared that speech had to be fitted in the frame of regular hand beats.

1. INTRODUCTION
L’objet ciblé par notre travail définit une recherche qui puisse observer opérationnellement la coordination d’un geste de percussion avec cet autre geste audible qu’est la parole.


Ce geste du bras produit, en l’occurrence, une séquence de percussions perpendiculaires lancées diffuse [5]; nous l’appellerons "geste de voile", comme la voile du marteau.

Décrit la coordination rythmique entre l’émission de la formulette et la production du batttement, par l’enregistrement de l’image et du son, tel est le but premier de cette communication.

2. CINÉMATIQUE DU GESTE DE VOILE
Le sujet (P.M., âgé de 65 ans en mai 1988, lors de l’enregistrement, chez lui à Autrans, Isère) a été filmé en extérieurs, en vidéo 8 mm (PAL), avec une caméra SONY CCD-V200. La posture de base utilisée, lors de l’effectuation du mouvement, est une position assise courbée [8] (Fig. 1). Le rameau de frène, reposant sur la cuisse, est tourné graduellement par la main gauche; seul le membre maniant le couteau se déplace, mettant en jeu deux segments corporels mobiles, la main et l’avant-bras.

Nous avons analysé ce mouvement de la main droite, via la caméra. Dans la présente description, nous n’avons retenu que 4 points significatifs (sur 7, cf. Fig. 1):

(a) Articulation métacarpo-phalangienne de l’index;
(b) Intersection lame-droit;
(c) Intersection lame-viole du couteau;
(d) Moelle du rameau de frène.

Un poste de numérisation et de traitement d’images [4] nous a permis de mesurer différents paramètres cinématiques, nous donnant trajectoires et fonctions temporelles, échantillonnées à 50 Hz.

Les paramètres retenus ici pour décrire les relations main-couteau et couteau-sifflet sont l’angle phalange-lame et la distance viole-moelle. Ces paramètres, édits en fonction du temps (Fig. 2), ont rendu possible le dépétage de plusieurs relations de phasage: la distance diminue à mesure que la valeur de l’angle augmente; et elle atteint son minimum à la première inflexion de variation angulaire, qui correspond à l’impact de la percussion (cf. zoom Fig. 2).

Ainsi l’organisation temporelle du cycle de voile (d’une durée de 260 ms, en moyenne) peut déjà se lire, sur le seul signal de la variation angulaire, comme un geste en trois phases (Fig. 2):

- lancé (depuis la flexion maximale jusqu’à l’inflexion de percussion);
- percédé (depuis cette inflexion jusqu’à l’extension maximale);
- relevé (depuis l’extension maximale jusqu’à la flexion maximale).

Ces trois phases ont respectivement, une durée moyenne de 80, 60 et 120 ms, soit 31, 23 et 45 % du cycle. Les études de gestes traditionnels comparables sont rares. Une recherche ethnotechnologique, réalisée en Normandie [1], nous a permis d’observer un grand nombre de percussions, dont celle d’un boulisseur, qui assouplit le cuir avec son marteau rivot. Avec une durée moyenne de cycle de 234 ms, décomposable en trois phases – une descente (lancé), un contact (qui correspond à notre phase de percédé) et une montée (relevé), soit 32, 23 et 45 % du cycle --, son organisation temporelle est en fait rigoureusement identique à celle de notre battement du sifflet.

Ces gestes possèdent une phase effectuant le lancé rapide (30%) et font donc partie d’une sous-classe de mouvements diadochokinétiques, la percussion impliquant une forte asymétrie temporelle.

3. ORGANISATION TEMPORELLE DE LA PERCUSSION EN FONCTION DU SIGNE, DE PAROLE
Sur le signal audio, échantillonné à 16 KHz, la mesure du cycle de percussion se précise, confirmant sa régularité : la variation angulaire de la moyenne (260 ms) est seulement de 30 ms (mesures prises sur le pic d’intensité); sur un nombre de battements donnant effectivement lieu à percussion (ce qui n’est pas le cas de certains battements "de démarrage", cf. infra), qui est exactement de 43 à chaque récital de la formulette.

L’étude de la relation temporelle entre le pic de percussion et le début de la voile suivante (c’est-à-dire l’établissement d’une structure formantuelle définie) a fait apparaître une variation importante, de 0 à 100 ms. Lorsqu’on examine la distribution de ces perceptions, on constate pourtant que celles-ci ne se produisent jamais avant la fin des voile précé dentes. Il semble donc que la contrainte de couplage impose que chaque percussion tombe au

minimum dans la phase obstruante du signal de parole, c’est-à-dire dans la phase qui est typiquement celle des consonnes.

4. CONCLUSION ET PERSPECTIVES
L’analyse de la performance de P.M. nous a permis de mettre en évidence une coordination – apprise dans l’enfance – entre geste et parole.

Les résultats obtenus révèlent un calage réciproque de la parole et du geste. Dans le démarre des séquences, le geste se calcule d’abord sur la parole: ce que révèlent certaines coupes donnés "à vue". Puis celle-ci doit s’ajuster dans le cadre d’une parfaite succession des battements: quelle que soit la durée intrinsèque des syllabes, chaque percussion doit tomber entre les voile, autrement dit “sur” les consonnes, en fonction d’attaque dans ces syllabes.

Nous sommes encore loin de comprendre suffisamment cette coordination geste-parole. La connaissance des "fréquences propres" des systèmes en jeu nous permettant de comparer que la fréquence d’ouverture et de fermeture du tractus vocal – qui correspond au rythme syllabique régulé par la main (cf. infra) – peut ainsi s’ajuster à la fréquence des battements régulés par le couple main-bras (qui est de 4-6 Hz en cadence râpide [7]). Cette coordination du geste et de la parole semble donc ici ralentir la fréquence de modulation du conduit vocal, puisque celle-ci est extradiée à 4 syllabes/seconde, par la cadence choisie pour le bras.

Cette première analyse devrait pouvoir nous informer, entre autres, dans ses développements, sur le paradigme illustré par Klapp [3]. L’une des "deux choses faites à la fois" étant la parole, il ne serait pas sans intérêt de tester la perception de la position des percussions dans la syllabe, par rapport à la théorie des Perceptual-Centers [6]. C’est ce que d’autres enquêtes et d’autres expériences devraient nous permettre d’aborder.

Cette recherche a été rendue possible grâce au soutien de l’Institut Dauphinois et du PPSH Rhône-Alpes n°20.
5. RÉFÉRENCES

ANNEXE
[tɔva 'sava si viʒ]  
meta paja meta ʁɛ  
sav'ec te pa si ʁɛ  
ka la 'merda do poʒɛ  
vɛ̃ ʁɛ  
di vɛ  
eklapa ʁɛ]  

Steve! Steve! Saint Vincent!  
Moitié paille, moitié foin.  
"Sèvrette" pas si bien  
Que la merde du poulin  
Viens bien!  
Dis rien!  
Fends "rien"!

(P.M., 65 ans, Autrans, 12-5-88; formulette du siflet, en dialecte provençal)
INTEGRATION OF AUDITORY AND VISUAL COMPONENTS OF ARTICULATORY INFORMATION IN THE HUMAN BRAIN

Riilo Aulanko and Mikko Sams

Department of Phonetics, University of Helsinki, Finland
"Low Temperature Laboratory, Helsinki University of Technology, Finland

ABSTRACT

In normal face-to-face conversation, both auditory and visual cues are used in speech perception. When the cues are contradictory, a perceptual "fusion" may arise, as in the "McCurk effect". Using magnetoencephalography (MEG), we measured the neural responses elicited by concordant and discordant audio-visual articulatory cues in the human brain. The auditory syllable [pa] was repeatedly presented to 10 subjects, together with a visual face articulating either [pa] or [ka]. The audiovisual stimulus, presented with different visual face stimuli, elicited different magnetic responses in the auditory cortex. This indicates that visual articulatory information has an effect on the processing of auditory phonetic information in the auditory cortex.

1. INTRODUCTION

Speech perception is audio-visual in normal face-to-face conversation. Seeing the articulatory movements of a speaker's face provides complementary information for speech comprehension. The visual cues are especially needed in a noisy environment and by listeners with hearing defects [1, 3, 15].

Visual information is obviously helpful, e.g., in discriminating between labial and non-labial consonant articulations or between rounded and unrounded vowels, but other distinctions are also reflected in the muscular movements of the face [7]. Even the difference between falling and rising intonation can perhaps be conveyed visually [3].

Visual articulatory information affects the perception of an auditory speech stimulus although people with normal hearing are not usually aware of this. A convincing example of the importance of visual cues is the illusion sometimes called the "McCurk effect". It refers to the phenomenon where a subject is presented with conflicting articulatory information through the auditory and visual modalities causing him/her to perceive speech sounds which are combinations or fusions of the visual and auditory cues [8-11].

The most frequently cited classical example of this audio-visual illusion is the case of an auditory syllable [ba] presented with a videotaped face articulating [va], eliciting an auditory perception of [da] [8, 9]. This illusion usually remains stable even after the subject is told about its nature.

There is no exact information about the actual neural basis of audio-visual speech perception. It has been stated that, after its preliminary analysis in the occipital cortex, the visual language reaches the angular gyrus where it is reorganized into auditory form [5]. It has also been pointed out that brain damage can impair the ability to lip read if a portion of the left occipito-temporal cortex [2].

In this experiment [13] we made neuromagnetic measurements to locate the neuroanatomical area in which the integration of auditory and visual components takes place. As a first step towards this goal, we wanted to see if visual articulatory stimuli have an effect on the processing of an auditory phonetic stimulus in the human auditory cortex.

2. EXPERIMENT

2.1. Subjects

Ten healthy adults (4 females, 6 males; 9 native speakers of Finnish, one of Swedish) were studied individually.

2.2. Stimuli

The stimuli were edited from a video recording of a Finnish female speaker articulating the CV syllables [pa] and [ka]. The auditory [pa] syllable was dubbed to the visual [ka] articulation, and combinations where the visual and auditory stimuli were in concordance (V=A, 84% of the stimuli) and combinations where they were discordant (V+Α, 16% of the stimuli) were joined to a continuous film of a speaker articulating one or the other of the syllables 800 times with an inter-stimulus interval of about one second. In seven subjects, the probabilities of the audio-visual stimuli were also reversed (V+Α 84%, V=A 16%). The auditory stimulus always remained the same syllable [pa] with a duration of 215 ms and an intensity of about 70 dB SPL. In a control condition, the face was replaced by a short green (84%) or red (16%) light (LED) stimulus, which preceded the auditory syllable by 350 ms.

2.3. Magnetoencephalography

The magnetic responses elicited by the stimulation were measured using magnetoencephalographic (MEG) recordings. MEG provides a powerful, completely noninvasive tool to investigate cortical activity in human subjects. In this method, the weak magnetic signals associated with neural currents are recorded outside the head by means of SQUID (Superconducting QUantum Interference Device) magnetometers [6]. The field is measured at several locations and its cerebral source is often modelled with an equivalent current dipole (ECD). The parameters of the model are the location, orientation, and strength of the source.

2.4. Procedure

During the experiment, the subject was lying on a bed in a magnetically shielded room with his head firmly supported, and the auditory stimuli were led to his right ear while he was watching the video monitor through a 12-cm diameter hole in the wall. In the control condition, the LED was attached to the wall beside the hole. The task of the subject was to listen carefully to what the speaker was saying and to count silently the number of all auditory stimuli, and to report the count after the session. Thus, the subject was not asked to react differently to the two stimuli. The only difference in reactions was supposed to be the different "stimulation condition". We could not ask the actual perceptual identity of each of the 800 stimuli from the subject during the experiment, but before the experiment we checked that the subject really heard the identical acoustic stimuli as two different syllables.

Magnetic field maps were constructed on the basis of recording over the left hemisphere with a 24-channel SQUID-gradiometer which possesses two derivatives of the radial component of the magnetic field at 12 locations simultaneously. The instrument detects the largest signal just above a dipolar current source. The exact locations and orientations of the gradiometers with respect to the head were determined by passing a current through three small coils, fixed on the scalp, and by analyzing the magnetic field thus produced.

The experiment consisted of presenting a frequent "standard" stimulus and an infrequent "deviant" stimulus in a pseudo-random order. In such conditions, an automatic neural detection process has been observed, the so-called mismatch response, which indicates that the nervous system has detected a change or difference in the repeated stimulation [12, 14].

3. RESULTS

The subjects perceived a strong audio-visual illusion: they heard the V+Α stimuli either as [ka] or [ka] or something in between.

The magnetic responses to the frequent V+Α stimuli typically consisted of three consecutive deflections, peaking at 50, 100, and 200 ms (Fig. 1). Similar deflections are elicited by speech sounds and can be explained by equivalent current dipoles in the supratemporal auditory cortex. The magnetic responses to infrequent V+Α stimuli had 50-ms and 100-ms deflections similar to those elicited by the V=A stimuli. However, starting at approximately 180 ms, the two responses were different. A rather similar difference waveform (responses to the frequent stimuli subtracted from those to the different ones) was obtained for infrequent V=A stimuli among frequent V+Α stimuli. However, the signals to the auditory syllables preceded by frequent green and infrequent red light stimuli were identical (Fig. 1).
The infrequent VeA stimuli elicited a distinct difference waveform in 7 out of the 10 subjects. The VeA stimuli elicited such a waveform in 6 out of 7 subjects studied, including those three who did not show it to infrequent VeA stimuli. Visual articulation presented alone, without the auditory input, elicited no response over the left temporal area in the two subjects studied.

4. DISCUSSION

The results of this experiment indicate that visual articulatory information has an effect on the processing of the auditory phonetic information in the human brain. Identical auditory syllables, presented with two different visual face stimuli, were heard as two different syllables. The neuromagnetic responses to acoustically identical but perceptually different auditory stimuli suggest that the processing of speech sounds in the human auditory cortex can be affected by visual input. The neural activity originating from the auditory cortex was not correlated with acoustical energy but with auditory, especially phonetic, perception.

The response distributions in this experiment could be explained by ECDS at the supratemporal auditory cortex, showing that visual information from the articulatory movements may have an entry into the human auditory cortex. This is consistent with the very vivid nature of the auditory illusion. We did not see coherent activity in the two areas suggested by Geschwind [5] and Campbell [2], i.e. angular gyrus and occipito-temporal cortex.

In face-to-face communication speech can be "seen" before it is heard. Visual cues from lip movements may exist in some cases hundreds of milliseconds before the corresponding auditory stimulus. Visual [ka] information might prime such auditory neurons which are tuned to any non-labial consonant followed by an open vowel. Due to priming, the auditory [pa] might activate the [ta] and [ka] "detectors" more vigorously than the [pa] detectors, giving rise to biased perception. Our control condition with light stimuli shows that the found difference waveform clearly cannot be explained by different degrees of attention allocated to the frequent and infrequent stimuli.

5. REFERENCES


---

FIGURE 1: Magnetic responses of one subject, measured with a 24-SQUID gradiometer over the left hemisphere in three measurement conditions. Only one of the channels with the largest responses is shown. The three pairs of traces were recorded over the same area in consecutive measurements. The number of averages is 500 for the frequent stimuli (84%) and 80 for the infrequent stimuli (16%). The recording passband was 0.05-100 Hz, and the responses have been digitally low-pass filtered at 60 Hz. The visually produced difference between the responses to the identical auditory stimulus can be clearly seen in the two uppermost pairs of traces. The responses to the auditory syllables preceded by frequent green and infrequent red light stimuli were identical (lowermost pair of traces).
AN OBJECTIVE AND A SUBJECTIVE APPROACH OF SPEAKER RECOGNITION

Elisabeth Lhote, Laura Abou Haidar

Laboratoire de Phonétique - 30 Rue Mégevand - 25000 Besançon - France

ABSTRACT
We consider speaker recognition as an integration level in the transfer process from production to understanding. In tackling speaker's recognition from the point of view of proximity between several speakers, we chose two complementary approaches: a "descending" approach, that allows extracting objective elements in both auditory and acoustic analysis, in order to associate voices unknown from the experimenter; a "rising" approach, that allows bringing to light objective criteria for the characterization of vocal proximity between speakers close at a genetic, acoustic or auditory level.

1. INTRODUCTION
Speaker recognition is considered here as a key process in speech recognition. The listener who recognizes someone by his voice resorts to various treatment mechanisms: for a global treatment, he refers to discourse analysis; for a local treatment, he selects acoustic characteristics which, memorized, become attributes characterizing one speaker. From the point of view of the listener, the two treatment models are associated and it is difficult to know whether one of them influences the other and how does the listener proceeds in distinguishing the two. It is often said that this approach is subjective. In fact, the recognition by the listener is done in real time: as soon as he hears the first words on the phone, he usually knows who is calling him amongst people he knows. This observation brings to the front, in daily practice, an ability to select and associate vocal attributes with a known person. However, sometimes, doubt disturbs recognition. The listener hesitates between two people. We are interested by this situation in as much as the listener's recognition system is not sufficient. We decided to tackle speaker's recognition from the point of view of proximity between several speakers.

2. HYPOTHESIS
Our hypothesis is the following: whatever the discourse of the speaker may be, and whatever his emotional state, the neuro-articulatory and neuro-phonicatory mechanisms which command and control the speech neurolinguistic programming are constant. This does not mean that the way we produce a syllable remains the same for each speaker, but that a neurolinguistic invariability remains as long as a pathological affection does not alter the voice.

3. EXPERIMENTATION
The experimentation focuses on comparison between different speakers, according to two complementary approaches: one called "descending", the other "rising".

In the first approach, we tried to associate unknown voices that had been recorded, with models. This "descending" approach allowed us on the one hand to extract objective elements in both auditory and acoustic analysis; on the other hand, we were better able to estimate the notion of proximity between voices.

In the "rising" approach, we selected speakers close in age, with family ties, with similar ways of talking, and having voices which are similarly confused on the phone. Then we tried to bring to light objective criteria allowing to characterize vocal proximity.

3.1. Descending approach
The first group was constituted of five speakers: S.A, S.B, S.C, S.D, S.E, and the second of twelve, among whom could be found the five speakers of the first group. In this case, we had to match voices of speakers reading a text varying between 2 to 5 minutes, of which only some sentences were produced by speakers belonging to both groups. The auditory analysis consisted of a systematic analysis of discourses at a phonetic level.

3.1.1. Global parameters
The global parameters which were the most pertinent were rhythm and intonation. In order to better bring them to light, we performed a simultaneous auditory analysis of two voices producing for instance the same sentences. A correlation between auditory and acoustic analysis allowed us to bring to the fore ways of speaking that are close and distant (FIGURES 1 & 2).

3.1.2. Local parameters
Afterwards, local analysis parameters were extracted by spectral analysis: a systematic analysis of formant trajectories in key sequences allowed us to put together or to separate some speakers (FIGURES 3 & 4).

The final results obtained with the help of this double analysis: local and global, auditory and acoustic, are positive and show the efficiency of this approach in discovering unknown links between voices and speakers.

3.2. Rising approach
In this case, speakers are known by the experimenter. The corpus is elaborated in order to bring to the light formant structures visible in key words or key syllables.

3.2.1. Acoustic proximity
Thirteen speakers produced the following text twice: "Tu sais, pendant les vacances à la montagne avec Jean, il y avait de ces tourbillons! Les tourbillons étaient trop forts!"

The selected syllable was [jõ] in "tourbillons". The results of this analysis showed a greater or lesser variability of slopes depending on the speaker. And particularly they allowed us to select 2 speakers whose slopes were very close. We recorded these two speakers again, and we asked them to vary their voice.

One sentence: "Les tourbillons de Lyon" was produced 40 times by each of them: 10 times in a normal voice, 10 times whispering, 10 times shouting, 10 questioning. We tried to extract a cue characterizing either the articulatory movement or an articulatory invariability.

The slope analysis of the two syllables [jõ] in different voices did not permit differentiation between the two speakers.

We noticed that the following cue: [F4 - F3] could be dependent of speaker's vocal behaviour: when converting these frequent values in tones, we noticed that this tonal cue seems to be an element that could characterize speakers' vocal behaviour:
* in the first speaker, the value of this tonal cue was 3 tones, whichever voice was used;
* in the second speaker, a variation of this cue was situated between two and three tones depending on the type of voice.

It is important to underline that from an auditory point of view, these two speakers don't have the same voice, even if the acoustic analysis shows a very close proximity.

3.2.2. Genetic proximity
We analysed three sisters' voices Y, L, N, two of which are often mixed up on the phone (L & N). We tried to find whether acoustic cues linked to formant
transitions gave an explanation of this proximity.
The tested sentence was the following: 
"Il y avait de ces tourbillons! Les tourbillons étaient trop forts!"

The key syllable was [i] in "tourbillons". We selected the slope of F2 between [i] and [o] and calculated it into tones. We think that this cue should contribute to define the velocity of the articulatory movement. We obtained the following results (FIGURE 5):

Number of tones for a 40 ms interval:
L → 6 tones
N → 4 1/2 tones
Y → 4 1/2 tones

Other experiments showed us that this tonal slope cue of the first three formants can be steady in some speakers production and unstable in others when they change from normal voice to shouting, whispering, questioning. We were expecting to find the same slope values in L & N, who are often mixed up on the phone; in fact, we didn't. We deduce from this result that results obtained at the auditory level can be different from those obtained at the acoustic level.

4. CONCLUSION
After having tested the relation existing between the auditory appreciation of a voice and its acoustic analysis - global and local -, we extracted the following points:
- Two voices auditorily close can be distant acoustically and vice versa; that is why it is important to associate the two approaches which should be considered complementary.
- If we are looking to characterize the articulatory movement velocity, it is useful to take into account the formant 4 and to use slope tonal variations.
- However, it should be noted that what appears to be necessary - during the rising approach - to the differentiation between two speakers is not necessarily sufficient to succeed in identifying a speaker from others during a descending approach.

In speakers recognition, as well as in speech recognition, a systematic correlation between the different analysis levels is necessary, in order to avoid favoring cues which belong to a unique analysis level.

5. REFERENCES

6. FIGURES

FIGURE 1 - Two Speakers close at the rhythmic and melodic level
Sentence : "C'est d'accord ou quoi?"

FIGURE 2 - Two Speakers distant at the rhythmic and melodic level
Sentence : "C'est d'accord ou quoi?"

FIGURE 3 - Formant trajectories of two close speakers

FIGURE 4 - Formant trajectories of two distant speakers

FIGURE 5 - F2 transition slope in the syllable [i]
FREQUENCY MODULATION OF FORMANT-LIKE SPECTRAL PEAKS

Peter J. Bailey and Kim Bevan

Department of Psychology, University of York, York YO1 5DD, U.K.

ABSTRACT

We report the results of two experiments showing that sinusoidal modulation of the centre frequency of one of a pair of formant-like spectral peaks increases its discriminability, as measured by the difference limen for spectral peak frequency. The apparent release from upward spread of masking afforded by modulation occurs for both noise-excited and pulse-excited stimuli and is not closely dependent on stimulus duration, modulation rate or peak centre frequency.

1. INTRODUCTION

The specification of formant frequency in vowel perception requires at least two potentially distinct stages: one logically-prior that isolates a spectral region corresponding to a local energy peak, and another that estimates the peak frequency. Errors are likely in the first step of identifying where formants are when spectral peaks are close in frequency, or when listening in noise, amongst competing sounds or with an impaired auditory system. Such errors will lead in turn to inescapable errors in the second step of formant frequency assignment and thus to probable inaccuracies in speech recognition performance.

Similar problems attend the visual perception of objects in complex scenes, where errors in locating the contours of an object can lead to conspicuous failures of visual identification. One powerful source of disambiguation in visual scenes is movement of the object or observer, which can provide cues to the appropriate parsing of the scene into figure and ground. In essence, the experiments reported here attempted to explore the utility of auditory object movement (an auditory object consisting of a single resonance) as a way of specifying for the listener the perceptual coherence of the energy contributing to a spectral peak. We hoped by this means to improve the accuracy of discrimination or recognition tasks that rely on the precision of the representation of peak frequency. We simulated auditory object movement using simple periodic modulation of resonance frequency.

There are demonstrations of the potentially beneficial role of modulation for both auditory detection and segregation tasks. Rasch [2] measured the masked threshold of a harmonic complex tone when it was mixed with a second harmonic complex of lower fundamental frequency. A 5 Hz, 4% vibrato imposed on the fundamental of the higher complex reduced its masked threshold by 17.5 dB relative to its threshold when unmodulated. McAdams [1] has shown that modulation of the fundamental frequency of one of a set of three concurrent vowels can increase judgements of its perceived prominence. Our experiments were concerned not with fundamental frequency modulation, but with modulation of spectrum envelope characteristics. In particular we sought to establish whether peak frequency modulation can enhance the discriminability of a spectral peak when presented against the background of an otherwise unmodulated spectrum envelope.

2. GENERAL METHOD

Our basic strategy for measuring the perceptual effects of frequency modulation involved four stimuli in each experimental condition. Two of the stimuli had a single spectral peak (the 'target' peak). In one case the peak centre frequency was not modulated and in the other it was sinusoidally modulated. The other two stimuli were like these, with the addition of a second lower-frequency spectral peak. In these two-peak stimuli the lower peak was never modulated and was sufficiently close in frequency to the higher-frequency target peak to impair unmodulated target peak discriminability. For each stimulus we measured subjects' difference limen (DL) for an increase in target peak centre frequency.

2.1 Stimuli

Stimuli were generated by passing broadband noise (experiment 1) or a 100 Hz pulse train (experiment 2) through digital second-order resonators. When two spectral peaks were required the outputs of two parallel resonators were summed. Resonator half-power bandwidths were fixed at 100 Hz (target peak) and 80 Hz (lower-frequency peak).

Fiber coefficients were updated at a rate of 1 kHz (experiment 1) or 200 Hz (experiment 2). The depth of modulation (i.e. the total frequency excursion) for the modulated peak was 16% of the centre frequency. All spectral peaks had approximately equal spectrum level (+2dB). Stimuli were presented at 70 dB SPL in broadband background noise at a level set for each subject to give the spectral peaks a presentation level of 10 dB SL.

2.2 Procedure

Difference limens were estimated using a two-alternative forced choice trial structure with two pairs of stimuli per trial. In one pair the stimuli were identical and in the other they differed in target peak frequency. The subjects' task was to identify the pair containing the different stimuli. The target peak frequency DL was taken to be the frequency difference corresponding to the '50%' response point on the psychometric function, determined by an adaptive staircase. Feedback was given after every response. Subjects were well practised before data collection began.

3. EXPERIMENT 1

In addition to the basic question of whether modulation of target peak frequency could improve its discriminability, the first experiment also explored the importance of modulation rate and stimulus duration.

3.1 Stimuli and Procedure

Target peak centre frequency was set to 1500 Hz. Target peak frequency DLs were measured for single-peak stimuli, and for two-peak stimuli with a lower-frequency peak at 1300 Hz. Since our major concern here was with modulation of spectrum envelope characteristics the resonators were not excited, producing whisper-like stimuli with relatively fully-specified spectrum envelopes.
Other stimulus manipulations were as follows. Modulation rate: (i) 0 Hz (unmodulated), (ii) 5 Hz, and (iii) 10 Hz. Stimulus duration: (i) 250 msec or (ii) 500 msec. Data were collected from seven subjects, including the second author.

3.2 Results
Mean DLs for all subjects are shown in Table 1.

<table>
<thead>
<tr>
<th>TABLE 1: mean DLs and standard errors (Hz) for experiment 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>modul:</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>250 ms</td>
</tr>
<tr>
<td>1 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
<tr>
<td>2 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
<tr>
<td>500 ms</td>
</tr>
<tr>
<td>1 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
<tr>
<td>2 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
</tbody>
</table>

For stimuli with a single spectral peak modulation increased target peak DL. However, the effect of modulation in two-peak stimuli was to decrease the target peak DL relative to the unmodulated condition, that is to increase discriminability. This was true for both modulation rates and both stimulus durations. DLs were smaller for 500 msec stimuli, but there were no reliable interactions between the effects of modulation rate and duration.

3.3 Discussion
The results of this experiment show that sinusoidal modulation of peak center frequency can lead to reliable improvements in the discriminability of a spectral peak when that peak is presented in an unmodulated spectral context. The absence of any interaction between modulation rate and stimulus duration shows that the effect is not dependent on the number of modulation cycles. Modulation appears to render the target peak perceptually more salient and thus less susceptible to upward spread of masking from the lower peak. This occurs despite the tendency for modulation to spread excitation around the peak frequency in the excitation pattern. The similarity in DL for one-peak and two-peak modulated stimuli suggests that modulation enhances the target peak with substantial immunity from the masking effects of the lower peak. In terms of the two-stage sketch of formant perception given in the introduction, it may be that modulation, by providing additional information for perceptual grouping processes, increases the efficiency of the first stage, in which the spectral region corresponding to a spectral peak is identified. The second experiment sought to replicate and extend the generality of these results.

4. EXPERIMENT 2
This was concerned with the dependency of the modulation effect on type of resonance excitation and target peak frequency region.

4.1 Stimuli and Procedure
Target peak centre frequencies were set to 1500 Hz or 900 Hz, with lower-frequency peaks when present at 1300 Hz and 700 Hz, respectively. All stimuli were pulse-excited with a constant fundamental frequency of 100 Hz. Other stimulus manipulations were as before. DLs were measured in 4 subjects for each target peak frequency. Most of the subjects had also served in the first experiment.

4.2 Results
Mean DLs for all subjects are shown in Table 2 for target peak frequency 1500 Hz, and Table 3 for target peak frequency 900 Hz.

<table>
<thead>
<tr>
<th>TABLE 2: mean DLs and standard errors (Hz) for experiment 2 (Target Peak frequency 1500 Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>modul:</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>250 ms</td>
</tr>
<tr>
<td>1 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
<tr>
<td>2 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
<tr>
<td>500 ms</td>
</tr>
<tr>
<td>1 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
<tr>
<td>2 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
</tbody>
</table>

TABLE 3: mean DLs and standard errors (Hz) for experiment 2 (Target Peak frequency 900 Hz) |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>modul:</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>250 ms</td>
</tr>
<tr>
<td>1 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
<tr>
<td>2 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
<tr>
<td>500 ms</td>
</tr>
<tr>
<td>1 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
<tr>
<td>2 peak</td>
</tr>
<tr>
<td>sem</td>
</tr>
</tbody>
</table>

The patterns of results for the two target peak frequencies was somewhat different. For pulse-excited stimuli with target peak frequency at 1500 Hz the results were similar to those obtained in experiment 1 with noise-excited stimuli at the same target peak frequency; as before, modulation apparently gave substantial immunity from the masking effects of the lower-frequency peak. For pulse-excited stimuli with target peak frequency at 900 Hz, modulation had the effect of decreasing the magnitude of the DL for single-peak stimuli as well as two-peak stimuli, relative to the DLs in unmodulated stimuli. As before, longer-duration stimuli tended to have smaller DLs, but there was no interaction between modulation rate and stimulus duration.

4.3 Discussion
The similarity between results obtained at the 1500 Hz target peak frequency for pulse-excited stimuli and those from the first experiment for noise-excited stimuli suggests that the enhanced discriminability of modulation affects two-peak stimuli. This is true of the spectrum envelope itself and not from the acoustic detail underlying it. We have data to suggest that the effect is genuinely attributable to modulation per se and not to phasic release from masking as the modulated target peak frequency increases above its mean value. The origin of the differences between the results for 1500 Hz and 900 Hz target peaks is not clear. One speculative suggestion is that modulation of the 900 Hz target peak may lead to detectable modulation of excitation in a larger number of auditory filters.

5. GENERAL DISCUSSION
We are aware that our account of the perceptual mechanism by which frequency modulation has its effects is crude and requires refinement. We believe the data are consistent with a role for perceptual grouping processes in the coherence that modulation imposes on the spectral energy contributing to a spectral peak. We are assessing the practical implications of these results by exploring the effect of second formant frequency modulation on vowel recognition.

6. ACKNOWLEDGEMENTS
The financial assistance of the U.K.SERC is acknowledged.

7. REFERENCES
VISUAL PERCEPTION OF ANTICIPATORY ROUNDING DURING ACOUSTIC PAUSES : A CROSS-LANGUAGE STUDY


* Laboratoire de Psychologie Expérimentale, CNRS URA 665
** Institut de la Communication Parlée, CNRS URA 368
Grenoble, France

ABSTRACT

This paper deals with visual perception of anticipatory rounding in French vowel-to-vowel gestures during acoustic pauses. Visual identification was studied for French and Greek subjects. Our results show that: (i) rounding anticipation can be identified only by eye several centiseconds before any perceivable sound; (ii) when the pause tripled, visual anticipation doubled, i.e. temporal positions of phonemic visual boundaries were dependent upon the extent of articulatory anticipation; (iii) but the boundaries steepness (switching time) was not; (iii) the comparison between French and Greek subjects did not revealed significant differences in rounding anticipation capture.

1.INTRODUCTION

Several studies in speech production have investigated anticipatory vowel rounding (of which, [1] is the most outstanding for French), particularly through consonant clusters, in order to investigate a major motoric issue, serial ordering.

As an expert in visual speech perception, McGurk mentioned briefly an unpublished experiment [5], with a reaction-time paradigm: it would appear to demonstrate that this anticipatory gesture can be detected visually to identify CV syllables from lip movements, prior to their being perceived auditorily. More recently [2] found, for French [zi zi] syllables, that the anticipation of the rounding gesture was perceived visually by the subjects who were able to identify the [y] vowel before the end of the [i], whereas it was not detected auditorily as early.

We studied, for French stimuli, visual perception of such an anticipation in vowel-to-vowel gestures without intermediate consonants, using natural productions of acoustically silent pauses between the vowels. Such pauses have, of course, a prosodically signalling function. So it is not the prosodic stream which is acoustically (if not visually) interrupted, but segmental information, here rounding. Consequently the general issue to be tackled is: can this segmental flow be tracked from the optic signal only, when the acoustics are disrupted?

In this paper, two specific questions are focused on: (i) is there visual information capture of the second vowel stimulus, prior to its acoustic onset, and, if so, how long before?; (ii) is there a shift in the visual boundary for speakers of Greek – who do not have the [y] vowel in their phonological inventory – by comparison with native French subjects?

For lack of models strictly dedicated to the audio-visual perception of speech anticipation (in spite of [6]), we will use here the predictions of three current articulatory models [7] and transpose them to the visual level, in order to evaluate which processing the “eyes” perform on speaker’s labial gestures: (i) the look-ahead model [LA] predicts a maximal anticipatory span, i.e. as soon as the rounding movement is possible; (ii) for the time-locked model [TL], movement onset occurs at a fixed time before the acoustic onset of the rounded vowel, (iii) the two-stage or hybrid model [H] allows to describe lip protrusion gestures with two components, a gradual initial phase, which begins as soon as possible in a look-ahead fashion, and a more rapid second phase (its onset is a peak in acceleration), which is time-locked.
Below: corresponding protrusion gesture for the upper lip (P1).
The left dotted line indicates the acoustic offset of the [i] and the right one the acoustic onset of the [y].
2. METHOD

2.1. Corpus

Words (i.e. transitions) which were embedded in a carrier sentence: "Tu dis : UHII sie?" ([y t d i i # y i i z]), "you say : ... , where UHII is, by convention, an "Indian" trapehe and "tie" is a third person present of a nonsense verb "tie". ([y t d i i # i i z] is the control stimulus with HII as "Indian"). Each transition had to be produced following two different pausing instructions, a short [ #] (and a long one [#]). Each sentence was repeated 10 times thus giving 40 utterances which were ordered in random order.

2.2. Video recording

A Turkish male talker was filmed, at 50 frames/second, with simultaneous face and profile views, in a soundproof booth. Talker's lips were made up in blue : Chroma-key was connected to the output of the front camera so that the blue could be filtered out. The talker was facing forward in a real time order in order to realize a minimal outlies detection of the lip slit. The subject wore black sunglasses in order to eliminate the red 1000 W Halogen floodlight, a slide rule was fixed on the right side of the goggles to ensure adequate profile articulatory movements.

2.3. Selection of visual stimuli

Four utterances were selected among 40 possible duration measurements of all interpausal voices. They were chosen as representative of mean durations for the short pause (## = 160 ms) and the long one ([ = 300 ms).

2.4. Articulatory processing

For each digitized frame (512 x 512 pixels), eight articulatory parameters, describing front slit and lateral protrusion characteristics, were automatically extracted by image processing [4] and kinematics (velocity and acceleration) were obtained by a cable spline smoothing of position functions. Examination of traces of upper lip protrusion (P1) vs. time (one of the usually available parameter in others studies, for [i # y] and [i # y] trajectories, revealed movements profiles with two components, i.e. hybrid profiles. Nevertheless (as in [7]), peak acceleration was not time-locked, occurring about 120 ms before the acoustic onset of the [y] in [i # y] versus 200 ms in [i # y]. Movement onset was neither time-locked (as in [7]), since it occurred 260 ms before the acoustic onset of the [y] in [i # y].

2.5. Test procedure

For 13 images for short transitions and 28 images for the long ones, with 3 images before pause onset and after pause termination. We thus obtained a total of 32 stimuli which were presented in random order, with a shift of 5 images between each subject. At the beginning of the test, 4 extra images were proposed to familiarize with the task. The stimuli were displayed individually to each subject on a high resolution computer screen. The task was to decide whether the stimulus was an /i/ or a /y/. Subjects were encouraged to answer rapidly (within a few seconds) via a computer mouse.

2.6. Subjects

25 French and 24 Greek normal-hearing native speakers served as native subjects (their hearing and vision acuities were checked). A good subject identification of the /i/ vs. /y/ contrast was confirmed for all Greek subjects (mean score : 93.5%).

3. RESULTS

The identification functions - traced from [y] percent responses for each image - have a classical S-shape (fig. 1 & 2). Of course corresponding displayed steady state profiles ([ i i ] images were generally identified as [i] (above 80%). Subjects were able to identify correctly (at 100%) "targets" of the present study, i.e. images corresponding to the non-silent outcomes of [i]. Moreover, they were clearly able to capture anticipatory visual information on rounding (95% correct) to 120 ms before the acoustic onset of the vowel, being French or Greek.

3.1. Differences and similarities in visual boundaries

A quantitative comparison between identification functions was achieved by Probit Analysis [3]. First, this method allowed us to date the position of visual boundaries (corresponding to 50% [y] responses) with regard to the acoustic onsets, and to test the significance of time differences. In addition, it allowed us to test the parallelism between functions, that is the deplying information on the possible similarity in steepness between the boundaries. For our final 2 boundaries took place 90 ms before the acoustic onset of [y] for French subjects, and 80 ms for Greek.

For [i # y]: boundaries anticipated of 180 ms, for French, and 190 ms, for Greek.

There was a reliable difference (at p<0.01) between the two conditions ([i # y] and [i # y], with each language group; i.e. when the pause tripled, visual anticipation doubled). But while the temporal positions of phonemic visual boundaries were dependent upon the extent of anticipation in pronunciation, on the other hand, the temporal accuracy of these boundaries (i.e. that steepness estimated by functions gradients) did not depend on the anticipation and were sufficient to switch from [i] to [y] in all cases.

On these two points, there were no significant differences (p<0.01) between French and Greek subjects. Notice that the Greek had a rather fair competence in auditory identification of [i] vs. [y] but their [y] productions were usually biased towards /i/. The other way round, they could have read the "U" choice as [i]. In both cases ([i] or [y]) however, they did not capture significantly less rounding anticipation than the French did.

3.2. Visual perception of anticipation and articulatory models

The observed significant shifts in boundaries could be yielded by the subjectsly disregarded the prediction of a time-locked visual anticipation. In fact, our perceptual as our articulatory (c.f. 3.2.2.) data allows us to reject strong versions of both TL and H models: neither onsets nor peak accelerations are time-locked on our temporal functions. What about the LA model? It can be rejected on the basis of our visual data only : while the anticipatory gesture begins as early as possible, the subjects ignore visually this change in the early accelerated transitions (fig. 3 & 4). More precisely, it is the position of the visual identification onset (detected as the first peak of the second derivative of the smoothed function) which reveals itself synchronous with the acceleration peak of the protrusion gesture (with a limit discrepancy of 1 image [23 ms] between these two events).

4. CONCLUSION

Rounding anticipation in vowel production has proved to be reliably identifiable only by eye several centiseconds before any perceivably sound (up to 120 ms). These results are at least valuable for stopped images. They need additional research on movement processing in speech (especially for acceleration detection) and further elaboration of appropriate models; neither TL, nor H. Cross-language comparison did not reveal significant discrepancies in visuo-temperal boundaries, whether the rounding dimension was bound to the front-back contrast, as in Greek, or whether it was free, as in French [i] vs. [y]. Whether this result argues for a universal lipreading skill, remains of course an open test.

* Many thanks to J.L. Schwartz and W. Semiclais for their advice in Probit Analysis and to T. Benvenit for improving our English.

S. REFERENCES

OCCLUSIVE SILENCE DURATION OF VELAR STOP AND VOICING PERCEPTION FOR NORMAL AND HEARING-IMPAIRED SUBJECTS

Yves CAZALS and Lionel PALIS

Laboratoire d'Audiologie expérimentale, Inserm unité 229, Hôpital Pellegrin, 33076 Bordeaux, France.

ABSTRACT

Reduction of silence duration in an intervocalic voiceless stop consonant induces misperception of voicing. Psychoacoustic results suggest that temporal resolution could be at the origin of this phenomenon. In this study a high correlation was found between boundary of silence duration and of voiced murmure duration which supports this hypothesis. In addition this study shows that for some hearing-impaired subjects the time boundary for voicing misperception can be considerably greater than for normal hearing. Most of these subjects present a simple temporal shift with a normally steep change of perception. So for them adjustment of silent occlusion duration could be a beneficial acoustical processing.

1 - INTRODUCTION

The shortening of the duration of silence in an intervocalic voiceless stop consonant has been shown to induce a misperception of voicing in normally-hearing listeners (Lisker 1957). The time boundary for this effect is about 60 milliseconds for French as well as for English (Lisker 1957, Serniclaes 1973, Lisker 1981). At the fastest speaking rates closure duration is about 60 milliseconds and on average occlusion time is shorter for voiceless than for voiced stop consonants (Lisker 1981, Port 1981). The misperception of voicing induced by shortening silence duration of an intervocalic voiceless plosive can be thought to be governed by classification of the shortest occlusive duration of silence as belonging to the voiced category. It can also be thought to originate from an insufficient delay for auditory excitation of the preceding vowel to decay. Results from psychoacoustical experiments on temporal resolution indicate that at low frequencies around 100 Hz which correspond to the voicing frequencies of adult males detection of a silent gap requires a gap duration of about 60 milliseconds (Shailer and Moore 1983, 1985, Green and Forrest 1989, Gorse et al. 1989). Several studies indicate that hearing-impaired persons show deterioration of temporal resolution (Fitzgibbon and Wightman 1982, Fitzgibbon and Gordon-Salant 1987, Glasberg et al. 1987, Nelson and Freyman 1987, Moore and Glasberg 1986, Gorse et al. 1989). It was reasoned that if decay of auditory excitation is indeed the basis for voicing misperception induced by shortening occlusive silence duration, some hearing-impaired individuals should show abnormal time boundaries for this effect.

Some previous studies dealt with temporal processing and the perception of stop consonants voicing for hearing-impaired persons. Voicing in initial plosives was found slightly reduced already (Paradis et al. 1981, Ginzel et al. 1982, Tyler et al. 1982, Johnson et al. 1984); more errors were found for final plosives (Revoile et al. 1982). And, two studies indicate that elderly persons require occlusive durations longer by about 10 milliseconds (Price and Simon 1984, Dorman et al. 1985).

This study investigated for the same hearing-impaired subjects voicing perception of an intervocalic voiceless plosive as a function of occlusive silence duration and also the degree of forward masking of the preceding vowel.

2 - MATERIALS AND METHODS

Twenty subjects participated in this study, eight normally-hearing and twelve hearing-impaired with a sensori-neural deafness.

Samples of natural speech tokens "aka" and "aga" were recorded from an adult male speaker. Speech waveforms were edited in a computer. From the "aka" sound eleven tokens were formed by varying occlusive silence duration from 0 to 200 milliseconds in steps of 20 milliseconds. From the "aga" sound one cycle of waveform during the voiced murmure was selected as having the same fundamental frequency as the "aka" sound. Bursts of murmure were then constituted by concatenations of this cycle and multiplication by a triangular envelope with a rise time of 20 milliseconds and a plateau adjusted from 0 to 180 milliseconds in 20 milliseconds steps. Ten final stimuli were made by altering these various bursts at the end of the first "a" of the "aka" sound thus constituting "a+voiced murmure" stimuli. These stimuli are meaningless to French listeners.

For tests all sounds were delivered monaurally through a Bayer DT 330 MKII headphone. Stimuli were presented at an intensity of 85 dB peak SPL at the maximum peak of the first a vowel. The contralateral ear received a broadband noise at about 85 dB above threshold. In a first test the various "aka" tokens were presented randomly ten times each and the subject was asked to respond each time by pressing a button marked "k" or "g" according to his perception. In the second test two stimuli were presented successively. The first was always the first "a" of the "aka" item and the second was one of the various "a+voiced murmure" token. Each "a+voiced murmure" was presented ten times randomly and the subject was asked to indicate whether the stimulus were different or not in anyway by pressing one of two response buttons. Before starting each test the subjects were familiarized with twenty to thirty presentations of the stimuli.

3 - RESULTS

Results from the first experiment are presented in figure 1. The score curves of identification of voicing as a function of occlusive silence duration for normally-hearing individuals were similar to those previously reported in the literature. The range of results obtained from normal listeners is indicated in figure as a shaded area. On the same figure all individual curves obtained from pathological ears are plotted. It can be seen that about one half of these curves lie within boundaries for normal ears, the other half exhibiting abnormal results. The curves outside the normal range all show, but in one case, a simple shift along the time axis keeping a steepness similar to normal curves.

![Figure 1](image-url)

Number of k responses

Occlusive silence (ms)

Results from the second experiment gave a series of curves having a similar shape as those of figure 1. Five hearing-impaired subjects indicated they could not perform this test in spite of some supplementary training.

From the score curves of experiments 1 and 2 the duration corresponding to a score of 75% was computed and served for analysis. A plot of the results of both tests is given in figure 2. It can be seen that for the normal ears the results seem to lie closely along a line, a correlation coefficient calculated on these data

---

54

55
4 - DISCUSSION

Results of this study show an abnormality long silence duration needed by hearing-impaired individuals to perceive correctly the voicelessness of an intervocalic velar plosive. Data from the second experiment support the idea that this originates from a deteriorated temporal resolution at voicing frequency.

The observed temporal shift in the hearing-impaired individual indicates that it may contribute to make their identification more vulnerable to fast speaking rates and to noisy background. This study revealed that the time shifts for hearing-impaired subjects are significantly longer than those reported for elderly persons in earlier studies (Price and Simon 1984, Dorman et al. 1985). The normal steepness of variations in perception may be a basis for improvements observed when speaking clearly for the hard of hearing (Ficheny 1986, 1989). It also indicates that such a signal processing could be useful to several hearing-impaired persons. The high correlation observed in this study between the first and the second experiment support the hypothesis of an abnormally long ringing at low frequencies after the cessation of a sound in some pathological ears. Other masking effects may also occur on the burst or formant transitions of the following vowel but they are quite unlikely since they would occur at higher frequencies where detection of temporal gaps requires much shorter durations (Shailer and Moore 1983, 1985, Green and Forrest 1989, Grose et al. 1989); the correlations with audiogram impairment at low frequencies also support this notion. The worst results associated with probable Ménière agree with physiological findings on experimental hydrops of altered coding of brief low frequency sounds (Cazal and Horner 1988).

Acknowledgements

The authors thank S Rosen, R Tyler and M Dorman for helpful comments. The participation of R Dauman MD, is gratefully acknowledged. This work was supported by a grant Grants-in-Aid.

5 - REFERENCES


PERCEPTION OF SYNCOPE IN NATIVE AND NON-NATIVE AMERICAN ENGLISH

Joann Fokes and Z.S. Bond

Ohio University, Athens, Ohio, U.S.A.

ABSTRACT

Native and non-native English speaking subjects made forced choice identifications of word triads embedded in phrases as spoken by three different English speakers. The triads consisted of 1) words with initial unstressed [æ] syllables, 2) words created by vowel syncope resulting in s-clusters, and 3) words containing s-clusters. A three-way analysis of variance revealed a significant interaction between the two subject groups, word triads, and the speakers. Native subjects were better able than the non-natives in identifying tokens even though there were no differential pacing in production. There was some bias in terms of speaker and particular word stimuli.

1. INTRODUCTION

Both native and non-native speakers alter the pronunciation of English in casual speech, but perhaps in different ways. For example, native Americans frequently employ syncope or vowel loss in the pronunciation of unstressed syllables. This phenomenon is well documented [3] in the case of internal unstressed syllables and appears to be correlated with word stress patterns. Such reductions seem to be more common in English than other languages because of its polysyllabic rhythm. Typically, syllables containing strong beats fall at irregular intervals and are surrounded or flanked by syllables with weak beats. Reductions also occur in initial unstressed syllables as in the casual pronunciation of 'spose for suppose. In fact, vowel syncope may spill over into the formal styles as in the network news commentary reporting recent 'Spree Court decisions.'

In the preceding example, vowel syncope results in a word with two juxtapositioned contexts resembling a dictionary word which does indeed contain a cluster. For example, vowel syncope as well as aspiration can be found in the production of sport which then becomes a possible homonym with sport. Just how listeners identify words containing vowel loss which become homonyms with real words is a question of interest in this investigation. It can be hypothesized that correct word identification is based on the semantic content of the message. On the other hand, there could be conclusions in the perception of the target word unless the phonetic characteristics of the utterance provide for cues in its correct perception. Thus, if the content is ambiguous, there could be phonetic information to aid in the perception of the intended word.

Before the perception of words containing vowel syncope can be adequately studied, the actual production of such items require description. The phonetic detail of clusters resulting from vowel syncope was previously investigated by Fokes and Bond [4,5]. They taped recorded ten American English speaking subjects and four non-native English speakers who read a series of six phrases or sentence sets. Each set contained a triad of test words embedded in the same phrase: 1) a word beginning with an unstressed syllable in the form of [æ] followed by [p] or [sp]; 2) a word containing an initial cluster consisting of [æ] or [sk] and 3) a word containing an artificially created [æ] or [sk] cluster resulting from vowel syncope. The subjects reported no difficulty in producing the syllables, which are real words, and naming such items as sport than the other members of the triad, sport and singleton. Each American listener could tell the phonetics of each phrase for all subjects were analyzed spectrographically. No group patterns were found for either American or non-native English speakers in their ability to differentiate real from artificial clusters in their speech. The stops in artificial clusters were not always aspirated. In addition, there was no show the expected systematic reduction in length of [s] in clusters as opposed to singletons reported in [3] and by Crystal and House [1,2]. Instead, individual subject patterns in the duration of the initial fricative, voice timing, or stop closure plus vowel were noted. Such individual patterns were not found among the non-natives. Rather, they lacked consistency within their own individual productions as if attempting productions in a trial and error approach. As expected, they also inserted vowels within the real clusters in which the Americans never did.

Since there were no consistent group patterns in the productions of subjects in differentiating words with unstressed syllables, real clusters or artificial clusters, one might predict that listeners would be unable to distinguish between the real and artificial clusters when embedded in the same phrase. Alternatively, if listeners are able to perceive artificial clusters as their target words with an unstressed initial syllable, there is likely to be a phonetic difference in the speech stream that was undetected in the studies by Fokes and Bond [4,5]. Of course, whether native listeners or non-native listeners also are capable of making distinctions resulting from vowel syncope.

2. METHOD

2.1. Materials

The stimuli for the present study were the productions from the previous investigation and consisted of two recorded readings of short phrase or sentence triads containing test words 1) with an initial unstressed vowel followed by [s], 2) a real [sp] or [sk] cluster, and 3) an artificial [sp] or [sk] clus-
ter. Each member of a triad was inserted into the following phrase sets:

On (sucumbing, scurrying, sicken)ing the party.

(secure, sked—ered, skured) the meat.

The (supplies, splice, sp'em) of tape.

My (support, sport, sp'ort) of baseball.

Four tokens of each item spoken by three native Americans and one proficient non-native speaker who had been speaking English since childhood were recorded in random order to make a listening tape of 192 items. The speakers were selected on the basis of clarity of the tape and the absence of any trace of an unstrained vowel. Words containing either the artificial or real clusters. The reduced vowel was present in the two words with the unstrained syllables.

2.2. Subjects

The two groups of subjects were 15 native American English listeners and 10 non-native listeners. The non-native group's experience with English was limited to academic training in English in their homeland and from two to five years' English contact at Ohio University. 2.3. Procedure

The subjects made forced-choice identifications (ex: splice/supplies) of each of the tape recorded tokens. Subjects listened via head-phones in a quiet listening laboratory.

3. RESULTS

The percent identifications of the triads by both groups of listeners are given in Table 1. The American listeners identified real clusters and two syllable words nearly 100% of the time. They heard the artificial clusters as two-syllable words at variable rates ranging from 86.4% for one of the native American productions to only 7% for the non-native proficient speaker.

Non-native listener identifications of real clusters ranged from 79% to 93% and from 56% to 93% for two-syllable words. They identified artificial clusters as two-syllable words from 18% for the non-native speaker to 57% of the native listeners. Interestingly, the non-native subjects perceived the proficient non-native speaker's artificial clusters as the target word more often than the native subjects.

Identification scores were also lexically dependent; 's_cumb was rarely heard as 's_cumb (8%), while 's_port and 's_spur were identified as two-syllable words 64% of the time. In fact, with the word 's_cumb removed from the analysis of test words, the artificial cluster rose to 59% for Speaker Four's productions and to 65% for Speaker Two. The latter also rose to a level of 38% for the non-native speaker productions as well.

Identification scores were submitted to a 2 by 3 by 4 repeated measure analysis of variance consisting of one between factor (two-syllable groups), and two within factors (4 English words and two triads). The Greenhouse-Geisser adjusted degrees of freedom were used to test the interaction and main effects. There were the following significant interactions: speaker by listener group (F = 4.74; df = 1, 23; p < .0001); speaker by word triad (F = 7.61; df = 3.26, 75.02; p < .0001) and speaker by word triad interaction. There was no group by word triad interaction. In determining the significant interactions, Speaker One was clearly different in that his artificial clusters could not be identified as intended by native subjects but were identified at somewhat higher rates by non-native subjects. Also significant were the main effects of listener group (F = 83.35; df = 1, 23; p < .0001); speaker (F = 45.97; df = 2.27, 52.11; p < .0001); and word triads (F = 64.22; df = 1, 44, 35.11, p < .0001).

4. CONCLUSIONS

The American native subjects were better able to identify artificial clusters as the target word contain ing untrained initial vowel than the non-natives. This ability cannot be credited to semantic cues only; the test words were embedded in the same phrase. Subjects, however, were highly influenced by specific words and the linguistic background of the speaker.

Because there was no single invariant acoustic pattern separating real from artificial clusters, we speculate that both groups of listeners were using multiple cues as a basis for perceptual judgments. That is, one speaker may have used a set of cues which, in turn, may have signaled the intended target word. Listeners may have the facility of adapting to the peculiarities of individual speakers and therefore may identify the target without apparent listening in a manner even when given a minimal amount of speech data.

5. REFERENCES


Table 1. Mean and 95% confidence intervals for native and non-native English subjects identifying the stimulus triads.

<table>
<thead>
<tr>
<th>Cluster Type</th>
<th>Native Mean</th>
<th>Non-native Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Clusters</td>
<td>95.2 ± 9.2</td>
<td>95.2 ± 9.2</td>
</tr>
<tr>
<td>2-Syllable Words</td>
<td>95.8 ± 9.6</td>
<td>95.8 ± 9.6</td>
</tr>
<tr>
<td>Artificial Clusters</td>
<td>95.7 ± 9.5</td>
<td>95.7 ± 9.5</td>
</tr>
</tbody>
</table>
ABSTRACT
Cerebral lateralization of speech processing depending on the type of the task presented, type of answering—vocal or manual, side of stimulation, etc. was examined. Dominance for different aspects of speech and complex non-speech sounds perception is shown. The paper presents the results of monaural testing in normal listeners the stimuli being amplitude-modulated noise and tones and CVC syllables with native and foreign vowels.

1. Introduction
Speech processing involves rapid decoding and construction of meaning from a transitory acoustic signal. The necessary linguistic processing is generally associated with the functions of the left hemisphere (LH). The last decades undoubtedly proved that predominant LH or RH involvement in speech processing – both perception and production. It was shown that LH mechanism provides for correct phonetic analysis, enabling to reduce sound continuum to functionally relevant segments, while the role of the RH is to realize global template recognition, discriminate the pitch, individual voice qualities, prosodic features. Our research shows that LH mechanisms secure accuracy of processing unfamiliar, novel material, while RH provides for quick orientation in familiar information. We have also shown the difference in hemispheric involvement in the perception and production of native and foreign languages. It is important to mention that both hemispheres can use various cognitive strategies depending on a number of factors including individual differences caused by genetically programmed lateralization of cognitive functions as well as those formed as a result of some specific training – language background including. Recent data exist that predominant LH or RH influence on information processing is determined by the task factor – the experimental or real and consequently the necessity of cognitive style choice: analytic for one class of tasks versus holistic Gestalt for the other. It is crucial that not all the stages of speech processing imply hemispheric involvement, i.e. higher cortical functions – lateralization can be the result of sensorimotor resolution capacities. This paper demonstrates the research in cerebral dominance for different types of information processing – perception, identification, imitation and categorization of speech and complex non-speech samples.

2. Methods
2.1. Experiment I
The subjects were 24 normal listeners between 20-50 years of age, all native speakers of Russian, right-handed. The stimuli consisted of CVC syllables which were CVC syllables made up of natural speech sounds produced by a male Russian-French bilingual. Russian stop consonants were used to construct syllables on a computer and record the stimuli. The resulting tape contained of 24 trials with 3-sec intervals. Each trial was presented individually. All possible combinations of hands and ears were used. The stimuli were asked to give single vocal manual response, to imitate the stimulus most accurately, to produce or write the Russian syllable similar to the target one.

2.2. Experiment II
49 normal subjects between 24 and 70 years of age were tested. The stimuli were amplitude-impulse-modulated sounds of different durations. Sounds were noise (frequency range 350-3000 Hz), sustained tones (250, 800, 1000 and 4000 Hz) and linearly frequency-modulated tones with rising and falling frequency changes (from 400 to 700 and from 700 to 300 Hz). The duration of a sequence of pulses was 0.08-3.2 sec., impulses being linearly rising or falling. The rhythm was 5-80 pulses per second (medium – 30 pulses per second). Subjects were asked to classify the stimuli according to two possible perceptual parameters – speech-like and moving. The stimuli were presented monaurally to the left and right ears in a quasirandom order. Subjects were instructed to respond monaurally (left or right). Reaction time was automatically registered.

3. Results
The subjects turned out to be grouped in two extremes the remaining arranged in between as to their psychophysical organization. The comparison of the group differences reveals (1) the "reciprocal" character of one of the groups: sharply different latent time depending on the stimulation sides, the parameters of the stimuli remaining the same and (2) the "synergic" group demonstrating ap-
proximately the same reaction time irrespective of the stimulation side and other conditions; subjects of this group make significantly less mistakes compared to those of the first one. Exploratory analysis reveals groups of subjects characterized by different hemispheric involvement in processing native and foreign language material - both vocal and manual reaction proved it definitely.

3.1 Experiment I

The data provided evidence of reaction time hierarchy in different task types. The first range is the time needed just to hear the stimulus and start reacting manually; the second - to decide which of the stimuli was presented and the third - to simulate articulation motoric of the stimulus without phonation. The greatest reaction time was registered when the stimulus was presented to the left ear, while the response was given by the left hand; the least - when the stimulus was presented to the right ear and the response was given by the right hand. It must be noted that though individual reaction times may vary around the measured value the relation between the ranges remains stable. Vocal responses also show hierarchy of latent times. It should be mentioned that processing of native versus foreign syllables seem to be controlled by different cerebral structures: "foreign" need mostly left hemisphere mechanisms - both for imitation and categorization; (probably it is caused by the necessity of phonemic coding), while native syllables can involve both (right and left) hemispheres.

3.2 Experiment II

The data showed three discrete ranges of stimulus durations revealed in classification tasks of amplitude-impulse-modulated targets according to their perceptual parameters: 0.08-0.2 sec.; 0.2-0.6 sec.; 0.6-3.2 sec. The subjects used these ranges to identify the stimulus as hoarse, speech-like (consonant-like with noise carrier and sonant-like with tone carrier) or moving in space (approaching with rising amplitude and moving away with falling one). It was shown that classification task is being solved within the same time limits irrespective of the stimulus acoustics, acoustic rhythm of pulses, duration, carrier frequency, amplitude shifting, the side of stimulation presentation, and the average-latent time was 1.5 sec. However, it should be emphasized that the usage of "speech-like" duration increases by 30 percent when the signal is being addressed to the right hemisphere. The findings suggest that classification procedure in the given experiment was based on dealing with individually formed functionally relevant template recognition. Opposite to it, experiments with amplitude changes identification show basic importance of (a) stimulus presentation side and (b) the use of the right versus left hand for the response. The maximum differences were examined in the range of "speech-like" durations revealed in classification experiment. The data demonstrate two main types of sensory-motor organization of speech: the dependence of lateralization on the experimental conditions - side of stimulation, type of stimulus (vocal/manual), ear/hand combinations, etc.

The results have basically revealed that classification and imitation procedures involve different hemisphere mechanisms depending on individual characteristics of subjects.

4. Conclusion

We put forward a suggestion that in central regulation of speech all high level processing of new and complex acoustic stimuli seems to be the function of LH, while familiar information engages both or RH preferably.

Speech processing therefore most probably uses higher levels in interpreting lower level of perception. LH provides for phonemic encoding and structural analysis of complex acoustic stimuli both in perception and imitation during short-term memory; RH realizes global template recognition. It should be emphasized that perception in language is specific and depends on individual acoustic and language background. The data demonstrate different types of organization of subjects irrespective of the type of experiment, which is impossible in interpreting mean or normalized data.

5. References


FACTORS AFFECTING THE GIVEN-NEW DISTINCTION IN SPEECH

Sarah Hawkins and Paul Warren

Department of Linguistics, Cambridge University, UK

ABSTRACT

Much attention has been paid to variation in acoustic properties depending on whether a word is "new" or "given" in a discourse. The hypothesis of this paper was that the given-new distinction is relatively unimportant in the perception of normal conversational speech. Selected words and CV fragments from those words were elicited from conversations with 3 people and their intelligibility was measured. Furthermore, the particular consonant involved and individual speaker characteristics all affect intelligibility more than the given-new distinction.

1. INTRODUCTION

Experiments show that the information a word contributes to discourse can affect its intelligibility: more predictable words tend to be spoken less clearly than less predictable words. Predictability that has been shown to affect intelligibility includes the meaning and grammar [6], and whether the word has been used before in the discourse [2]-the so-called new-old, or given-new, distinction.

These differences in intelligibility are statistical tendencies: not all words are affected, and some of the differences are small. Moreover, whereas some studies find differences in acoustic measurements that correlate with intelligibility differences, others find no differences in the same parameters, albeit in different languages [cf. 2; 4; 5].

If the given-new distinction has a significant influence on the intelligibility of all speech, there would be important consequences for models of both human and machine speech recognition. However, this paper reports preliminary work intended to investigate the possibility that "given versus new" is too simple a distinction to be useful for normal conversational speech.

One challenge in studying the given-new distinction is defining what is "old" information. Most studies treat the first instance of a word as new, and later instances as old. While this may be appropriate in an analysis of the discourse, it is unlikely to be appropriate for predicting the intelligibility of individual words or parts of words in ordinary conversations. A second or later word may be spoken in isolation, or with contrasting stress, for example, both of which might be expected to increase rather than reduce its intelligibility. We do not question that predictability is one factor that can affect intelligibility. But we do suggest that its normal conversational speech, the given-new distinction has only a small effect on intelligibility; other factors will be at least as influential.

Patterns of intelligibility are likely to depend on the types of discourse and speech being analysed. Large intelligibility effects due to the given-new distinction have tended to be found with speech that has been controlled for several aspects of linguistic context, or with tasks where clarity of speech and style of presentation are crucial [1; 2]; even here, intelligibility also varies with the information content of the repeated word and the experience of the speaker [1].

Fluent reading of texts may give a distorted view of the prevalence of given-

new distinctions in speech. Texts designed to elicit such differences in intelligibility are likely to produce them. But these differences may be much less likely to occur in normal conversational speech, which typically has shorter and less grammatically complex phrases. Hunnicutt's [4] finding that a greater intelligibility effect arises with long sentences typical of the written but not the spoken language supports this view.

Word intelligibility is also likely to be influenced by phonetic factors. The prosodic context has already been mentioned. Differences due to segmental-phonetic structure could depend on the acoustic properties of the sounds involved and/or to the phonological inventory of the particular language. For example, stridency is normally a robust acoustic property, and the range of possible articulations for a strident sound is fairly small. Thus stridency involves relatively little spectral variation even in casual speech. For languages in which a strident /a/ is contrastively distinct, and phonemically contrastive, then, strident sounds might be expected to retain a high level of intelligibility in most contexts.

A phonetic distinction that is mainly dependent on phonological space is the leniting of velar stops in English. The only English consonants are oral and nasal stops; so, since /h/ can only be syllable-final, and the acoustic correlates of nasalization are fairly distinctive and distributed over time, leniting /h/ and /h/ is unlikely to pose problems for the listener. In contrast, alveolar stops share a crowded section of English phonological space, and typically are not unlike strident fricatives in some of their spectral properties. In comparable phonetic environments, then, we would expect velar stops to vary more than alveolar stops in manner of articulation.

2. EXPERIMENT

To examine the worth of these arguments, we collected from natural conversational speech repeated tokens of the same words spoken by different people. We then measured the intelligibility of the whole words and their medial consonant. The words were all bisyllabic and stressed on the first syllable. The medial consonant was (a) the sound of interest (b) where the word became lexically unique, and (c) one of /d g s l/.

Medial consonants were chosen so that, as far as possible, the immediate phonetic context was controlled for coarticulation effects. Medials also allow the possibility of presenting CV, VC, and VCV portions of the words to listeners for identification. Requiring the medial consonant to represent the word's uniqueness point greatly constrained the choice of words, but had the advantage that word identification would take place under similar conditions of lexical access [cf. 7].

The choice of sounds was governed by the existence of suitable words and by the following considerations. 1. /s l/ are strident; the others are not. 2. /g/ will vary in manner of articulation more than the others, so under comparable conditions its intelligibility should vary most. 3. The experimental manipulations and acoustic analyses are more straightforward for voiced than for voiceless stops [3]. 4. The fricative /h/ resembles /h/ in that it is long (so could have an intelligibility advantage when excised from running speech), but it is nonstrident.

3. HYPOTHESES

Over the whole corpus:

1. First tokens of words and of medial consonants from sentences of new intelligibility from second tokens. This will also be true for the subset of first and later tokens bearing nuclear stress.

2. Tokens with nuclear stress will be more intelligible than with secondary or no stress, regardless of how many times the word has been used in the conversation.

Isolated sounds will differ in intelligibility such that:

3. Strident ([s l]) sounds will be more intelligible than other sounds, and nonstrident later instances will be as intelligible as the first instance.

4. Because we expect /h/ to vary more than /h/, /h/ will be likely to show variation due to the given-new contrast and to differences in sentence stress.

5. People will differ in the overall intelligibility of their speech and in how much it conforms to these predictions.
4. METHOD
The selected materials were sorted into four topics. Two women and one man, speakers of Southern British English, each repeated them with the experimenters in a sound-treated room. The speakers all knew the experimenters, and spoke in relaxed conversational style. Pictures were used to stimulate and guide discussion towards the words we were looking for. In the vast majority of cases the experimental subjects were the first users of the words of interest.

The repeated experimental words selected from within each speaker's discussion of the relevant topic were: 1. the first phrase; 2. the second production; 3. where possible, the next production contrasting in stress with the second token. In this paper, the third tokens are only used in comparisons of nuclear with other stress levels. The remaining 21 word sets were digitally excised from their fluent contexts and recorded onto digital audio tape for presentation to listeners.

For word identification, tokens were heard in white noise at a signal-to-noise ratio of 5 dB above the average intensity of the speech (excluding silence). Each subject heard only one token of each test word, counterbalanced across nine versions (3 speakers x 3 repetitions). The ISI was 4 ms, during which subjects wrote down the word they had just heard. Each test list had between 17 and 19 words and was preceded by 6 practice words.

In a second task, fragments containing consonantal information were excised: for stops, the burst and following 80 ms; for fricatives, the friction period plus 40 ms of the following periodicity. No noise was added. Each listener heard all excised segments in one of two randomisations, preceded by a 6-item practice list. The ISI was 2 ms, with a longer ISI after every tenth item. Listeners wrote down the consonant(s) they heard.

90 students completed the word identification task (10 on each version); 10 further students took part in the consonant task. Both tasks were open response. Listeners heard the materials over headphones in a sound-treated room.

5. RESULTS
The predictions were tested using ANOVAs, with designs differing according to the experimental hypotheses. Comparisons of the more interesting results so far. Differences reported as significant achieved a probability of 0.05 or better.

Words. Following [2,3] a response was scored as correct only if the whole word was identified correctly. Our argument that conversational speech should show no general tendency for the new-given distinction to appear is supported by the finding of no overall effect for this factor in the intelligibility scores. In contrast to this, we find a clear effect of stress type: words carrying nuclear stress are significantly clearer than others (68% vs 50%). Taken together with the distribution of stress types in our sample, this gives a long way towards accounting for the lack of a new-given distinction.

The new items almost all have nuclear stress (92%), and so do a large minority of the given (44%). Unsurprisingly, amongst the words carrying nuclear stress, there is no effect of new vs given. There were also no overall speaker differences for word intelligibility.

In an attempt to control for some of the variability in parameters other than that of new vs given, we chose a subset of materials with comparable phonetic makeup (one word, produced by all speakers, from each of the five sound types). In this subset new items are significantly more intelligible (78% vs 45%). However, it is possible that there is a context here of prosodic context, since 13 of the 15 new items are in nuclear position, but 4 of the 15 given. Further work is needed here.

Consonants. In scoring of the identification of consonants, we are interested primarily in place and manner; errors in voicing only are therefore counted as correct. As expected, we found significant effects of sounds and speakers. Subjects fricatives achieved by far the best scores /f/: 91%, /v/: 87%; with /θ/ and /ð/ intelligibility (56% and 55%) and /ʃ/ /ʒ/ worst (19%). The stress effect found for the word task is replicated here, with significantly fewer errors for consonants from words bearing nuclear stress (66% vs 53%).

6. CONCLUSION
Our hypotheses regarding whole words (1 & 2) were supported by the general finding that sentence stress affects intelligibility more than the simple given-new distinction. The hypotheses for consonants were partially supported in that strident fricatives were always highly intelligible (3), and in that given-new differences appeared for /θ/ but not /ð/ (4). However, the sentence stress effect found for whole words did not appear for isolated consonants. Whereas speakers' whole words did not differ in intelligibility, there were large differences in the intelligibility of their isolated consonants (5). This finding suggests that individuals vary in how much they distribute acoustic cues within words; listeners' perceptual strategies must show the required flexibility [cf. 3,7].

The /θ/ and /ð/ groups were evaluated further to compare the contrast in "stridency" and "phonological space" discussed above. The figure shows the predicted significant interaction of sound (/θ/ vs /ð/ with given-new, as well as main effects of speaker and given-new. On the whole, /θ/ loses intelligibility on repetition whereas /ð/ does not, but the effects are much greater for some speakers.

Figure: Consonant intelligibility for /θ/ vs /ð/ subject.
Different line styles denote the 3 speakers. Crosses show /θ/ and squares /ð/ identification scores.

7. REFERENCES

68

69
ABSTRACT

It is now well recognized that the right hemisphere is concerned with processing prosodic features of speech—intonation, rhythm and stress. There are however contradictory data concerning linguistic prosody as most of the research involves affective stimuli only. The present paper deals with neural aspects of both kinds of prosody in normal listeners. The results show hemispheric specialization for linguistic and affective prosody, the latter being a complex continuum.

1. INTRODUCTION

A role of the right hemisphere in the mediation of emotional speech was shown as early as 1874 by H. Jackson who observed that emotional words (i.e. ourises) were selectively spared in some groups of aphasics. In 1947 J. Martin-Krohn demarcated the processing of affective and linguistic prosody. He was one of the first to show right hemisphere dominance for emotional characteristics of speech. During the last twenty years a special role for the right hemisphere has been demonstrated for emotional processing, based on studies examining expression and understanding of emotion in brain-damaged patients and normal subjects. Nevertheless in the majority of papers comprehension and production of intonation as a whole is still being associated with the function of the right hemisphere, "intonation" interpreted by brain specialists as emotional characteristics of speech, linguistic intonation being neglected. There are a lot of contradictory data, showing not only right hemisphere, but left hemisphere involvement in processing intonations of different types. Some results are difficult to interpret because of the principle difference in investigation procedures, stimuli sets, types of questionnaires, etc. In fact there is no adequate hypothesis for laterality of any prosody yet. The present paper covers part of a cross-cultural investigation of hemispheric role of processing affective and linguistic prosody carried out in normal subjects and in brain-damaged patients. The aim of the study is to clarify the extent to which traditionally known right hemisphere involvement in the process is adequate.

The paper deals with neural representation for the perception and imitation in normal listeners.

2. METHOD

2.1. Subjects.
Male and female adults, postgraduates, aged 20-50, right-handed.

2.2. Stimuli.
The stimuli were Russian phrases of different prosodic types—both linguistic and affective. The set was formed of (i) communicatively different phrases, designating types distinguished from each other by intonation alone; (ii) syntactically different phrases—declarative, interrogative, imperative, exclamatory, etc. (iii) phrases with differing sentence accents, depicting semantic factors revealing communicative centers of the sentence—arbitrary syntactic complexity with meaning differentiating prosody; (iv) prosodic type, expressing surprise, politeness, anger, delight, etc., all chosen at random. The stimuli were read and recorded by a professional.

2.3. Procedure.

Every subject was listening to the same recording. The stimuli were presented monaurally to either the left or the right ear in random order. Noise being presented to the other ear. After the presentation of every sentence subjects were asked to choose one of the three answers printed on the test-cards. The reaction time and types of answers were registered.

3. RESULTS.
The data demonstrate right-hemisphere advantage for processing emotional stimuli—there were significantly fewer errors and the shortest latency period when the stimuli were presented to the left ear than to the right one. Communicationally, or syntactically different phrases appeared to be a complex perceptual domain—intonation types "analytical" seem to involve left hemisphere, while the others "Gestalt-like" show a privileged role of the right hemisphere. Sentences of different phrase accents showed surprising laterality effects—the majority of subjects revealed left-hemisphere dominance according to reaction time and correctness of answers. This stands in marked contrast to the results for prosody perception reported earlier. Adequate imitation of prosody did not reveal definite right hemisphere superiority as it could be expected a priori. It appeared that cognitive and communicational validity, the degree of syntactic complexity and novelty can produce strong effect on hemispheric preference.

4. CONCLUSIONS.

Our previous research demonstrated that right-hemisphere mechanisms may be responsible for adequate actual sentence division and for other semantic factors needed for sentence interpretation (e.g. prosodic expression of given/new distinction hemispherical sentence perspective). Our experiments in linguistic
competence show that cerebral hemispheres play essentially different roles: the right one operates largely with extralinguistic reality, it relates sign to its different. The left hemisphere interrelates signs, refines the process of speech production. In analyzing grammar it uses transformational rules while the right hemisphere uses "given/new" strategy, which in Russian may be provided by the definite word order of speech prosody - the fact that has never been investigated in the light of hemispheric specialization. The findings under discussion suggest that not only linguistic prosody may be associated with left hemisphere mechanisms versus right hemisphere mechanisms as emphatic but that linguistic prosody itself is most possibly divided between the hemispheres depending on the semantic factors.

In our study we find evidence for left-hemisphere preference for the linguistic types of prosody and right hemisphere preference for emotional prosody which is in accordance with literature data from brain-damaged patients. The most informative appeared to be sentences of different actual sentence division. The perception of such phrases demonstrated surprising laterality effects - the majority of subjects revealed left hemisphere dominance for complex phrases that needed special analysis versus right hemisphere dominance for wellknown, previously familiar "gestalt-like" phrases, psychologically "idiomatic".

We consider these findings to be of interest because of several factors: (i) normal subjects used for the procedure, (ii) linguisticly balanced stimuli, (iii) new type of procedure - notes for masking the other side of perception, reaction time measurement, specially designed "answer-cards", etc.

NOTE: The help of prof. N.Svetozarova, Leningrad State University, in tape construction and recording, and her invaluable comments are gratefully acknowledged.

5. REFERENCES.


REFERENCES.


MONRA-KROHN J.H. (1947), "Disprosody or altered melody of language", Brain, v. 70.


L'INFLUENCE DE LA DUREE DANS L'IDENTIFICATION DES LIQUIDES: ETUDE COMPAREE EN ESPAGNOL DE BUENOS AIRES ET EN FRANCAIS DE MONTREAL

Benjat Jacques*, Maria Amalia Garcia Jurado** et Miguelina Guiran**

*Université du Québec à Montréal, Canada, et **Laboratoire de Investigations Sensoriales, CONICET, Buenos Aires, Argentine

ABSTRACT

This paper compares acoustic and temporal cues of /l/ and /l/ in Montreal French and Buenos Aires Spanish with their identification in syllable context. Two Argentinian and two French Canadian speakers read short sentences in which /l/ and /l/ figure in CV, VC and /l/ CV contexts with the vowels /l/, /a/ and /l/. Segments of the waveform were chosen for perceptual analysis. It was found that, most of the time, both languages, the listeners perceived as modifications of the intensity of the timbre of the contiguous vowel and that they cannot be identified unless the selected segment contains three or more cycles of the vowel. The modifications take different shapes according to the language, the consonant, its place in the syllable and the contiguous vowel.

1. INTRODUCTION

Ce travail fait partie d’un projet plus vaste portant sur l’analyse des similitudes et des différences entre les consonnes latérales et vibrantes du parler espagnol de Buenos Aires et du parler français de Montréal. Des études antérieures ont présenté des observations sur les propriétés acoustiques et perceptuelles de ces consonnes en espagnol (Giraud et Rosso [4], Garcia Jurado, Guirro et Rosso [3]) et en français (Chatoffrout [1.2], Sinafra [5-6], Tournier [8]). Nous proposons de comparer les principales caractéristiques acoustiques et temporelles de /l/ et /l/ en relation avec leur identification en contexte syllabique, en particulier pour déterminer la durée minimale nécessaire pour l’identification de ces sons et d’analyser les changements qui interviennent dans la portion critique du segment temporel, c’est-à-dire la portion où il y a recouvrement des timbres de la consonne et de la voyelle adjacente. La présente étude doit être complétée par une étude perceptuelle faisant appel à des auditeurs des deux langues.

2. METHODE EXPERIMENTALE

Quatre locuteurs masculins, deux argentins et deux canadiens, ont enregistré des phrases courtes contenant les émissions de /l/ et /l/ en contextes de CV, VC et /l/ CV avec les voyelles /a/, /l/ et /l/. L’onde complexe observée par l’auditeur et dont des exemples sont illustrés dans les figures 1 et 2 de la page suivante a servi de base à l’étude acoustique. On pouvait y voir la voyelle, la liquide, ainsi que la portion critique. Un traitement de ces sons a été effectué en tenant compte de leur variation dans l’ordre temporel. Nous avons sélectionné au moyen de curseurs des segments du tiers d’unités de chacune des syllabes émises. Nous avons ensuite écouté la portion correspondant à la consonne afin de déterminer si elle pouvait être identifiée isolément. Puis les segments ont été amplifiés de leurs extrémités à commencer par celle de la voyelle jusqu’à l’obtention de la portion temporelle minimale nous permettant de percevoir encore la voyelle. L’étude a porté sur ce segment minimum.

3. RESULTATS

D’une façon générale, les liquides se présentent comme une modulation du timbre ou de l’amplitude de la voyelle adjacente. Cette modification peut s’étendre à toute la voyelle, ou se limiter à une partie de celle-ci. Lorsque seulement une partie de la syllabe est ainsi modifiée, il est possible d’observer un "trading off" ou une plus grande durée de la voyelle modifiée avec une durée plus brève de la voyelle libre. L’étude a porté sur ce segment minimum.

3.1. Syllabes avec /l/

3.1.1. Positions initiale vs finale

En espagnol, lorsque /l/ est en position initiale, on observe un segment de base amplitude dont le timbre est celui d’une voyelle neutre suivie d’un autre de plus grande amplitude qui correspond au moyen vocalique. La transition entre les deux segments est abrupte avec /l/, moins abrupte avec /l/ et graduale avec /l/. En français, le syllabe avec /l/ présente les mêmes caractéristiques que sa correspondante espagnole. Par contre, avec /l/, on a observé chez un locuteur une diplongaison, /l/ étant perçu comme un [l] et la syllabe, [i], au lieu de [i]. Enfin, lorsque /l/ est la voyelle, chez un des informateurs, il y a superposition complète entre celle-ci et la liquide, de sorte que le segment entier se perçoit comme un [l] prolongé.

En position finale, /l/ espagnol montre une transition graduelle entre les voyelles /a/ et /l/ et la liquide celle-ci se perçoit comme une modulation d’amplitude de celle-ci. En contexte de /l/, chez un informateur, il y a superposition complète de la liquide et de la voyelle, alors que chez l’autre, on perçoit d’abord un [l] suivi d’une pulsion sans timbre défini, elle-même suivie d’un autre [l] d’amplitude plus faible que le premier. En français, on observe une transition graduelle des deux voyelles avec la consonne, à ce qui que, dans les contextes de /l/ et /l/, s’ajoute une vitesse, /l/ et /l/ étant perçus respectivement [l] et [l]. À noter que le /l/ précédant /l/ dans les segments espagnol est la voyelle brève ouverte comme dans /bol/, non le /l/ fermé long de pollo. À l’exception des conures où la liquide et la voyelle se recouvrent complètement, la durée du segment critique pour les deux langues est de l’ordre de 20 à 30 ms en position initiale et près du double en position finale.

3.1.2 Position intervocalique /a/ /CV

En espagnol et en français, les séquences /a/ /l/ et /a/ /l/ reproduisent les mêmes phases que celles de l’espagnol observées dans les combinaisons /a/ /l/, /a/ /l/ et /a/ /l/ et les segments critiques sont de durée comparable en espagnol, légèrement supérieur en français. Quant à la séquence /a/ /l/ de l’espagnol, si on peut bien reconnaître le passage de la première voyelle à la liquide, il n’ay a pas aussi bien le passage de celle-ci à la consonne que, dès le début de son dénouement, /l/ prend le timbre de cette voyelle. Pour ce qui est de la même suite en français, on constate que /a/ /l/ se transforme en [e] par superposition partielle avec la voyelle /l/ et que la liquide prend elle aussi ce nouveau timbre /a/ /l/ devient [e].

3.1.3. Les durées minimales

TABLEAU 1

<table>
<thead>
<tr>
<th>Durées minimales des segments permettant de reconnaître la liquide /l/ (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informateurs</td>
</tr>
<tr>
<td>1er</td>
</tr>
<tr>
<td>/l/ /l/</td>
</tr>
<tr>
<td>/l/ /l/</td>
</tr>
<tr>
<td>/l/ /l/</td>
</tr>
<tr>
<td>/a/ /l/</td>
</tr>
<tr>
<td>/l/ /l/</td>
</tr>
<tr>
<td>/l/ /l/</td>
</tr>
<tr>
<td>/a/ /l/</td>
</tr>
<tr>
<td>/l/ /l/</td>
</tr>
<tr>
<td>/a/ /l/</td>
</tr>
</tbody>
</table>

*Superposition totale de la voyelle et de la liquide.
Les proportions de durée du tableau 1 occupées par la voyelle plus variante entre 350° et 660° pour la voyelle initiale et finale. Des exceptions s'observent dans le contexte de /ə/; cette voyelle occupe en effet 720° de la durée du segment espagnol de 170 msec et y a pour les sons complémentaires des deux sous en français la position finale, la proportion de voyelle libre se situe entre 25% et 45%. La durée minimale nécessaire pour identifier la liquide est donc plus grande lorsqu'elle est en position finale. En outre, pour chacune de ces deux positions, à une exception près, c'est le contexte de /ə/ qui montre les durées les plus longues. Toutefois, en position intermédiaire, il y a lieu d'observer dans les deux vocaliques avant et après la liquide parce que ces voyelles fournissent un appui favorisant la reconnaissance de celle-ci.

3.2. Syllabes avec /ə/

 généralement, /ə/ espagnol se réalise comme une interruption sur un silence dans le segment vocalique. Ceci est surtout vrai en position initiale de syllabe. En position finale, en effet, cette coarticulation peut parfois se réaliser comme une interruption d'un silence de qualité de la voyelle précédente. En français, /r/ peut présenter des vibrations gutturales faibles, observées dans les traits accoustiques, mais pas dans les traitements. Il peut aussi présenter des formants sans vibrations. Dans les deux cas, le résultat en plan perceptuel est une modulation de la voyelle adjacente. Il peut également se réaliser comme une fricative, à l'issue de /r/ perçue. Enfin, en finale, il peut également se réaliser comme une /r/ précédée.

3.2.2. Position intermédiaire /ə/CV

En espagnol, on voit se reproduire pour les suites /ər/ et /ər/ les mêmes phénomènes que dans les suites déjà observées de /əl/, /əl/ et /əl/. En ce qui concerne /əl/, la présence des deux voyelles adjacentes est nécessaire pour l'identification de la consonne; celle-ci est suivie d'un silence entre les deux voyelles; dans /əl/, il y a un silence de la seconde, mais dans /əl/, chez un informateur, il adopte le timbre de la première voyelle et, chez l'autre, il se trouve comme un /u/.

3.2.3. Les durées minimales

Durées minimales des segments permettant de reconnaître la voyelle /ə/ (msec)

<table>
<thead>
<tr>
<th>Informateurs</th>
<th>Argentina</th>
<th>Canada</th>
</tr>
</thead>
<tbody>
<tr>
<td>1er</td>
<td>2er</td>
<td>1er</td>
</tr>
<tr>
<td>/r/</td>
<td>90</td>
<td>73</td>
</tr>
<tr>
<td>/ə/</td>
<td>62</td>
<td>78</td>
</tr>
<tr>
<td>/ər/</td>
<td>95</td>
<td>114</td>
</tr>
<tr>
<td>/ər/</td>
<td>97</td>
<td>146</td>
</tr>
<tr>
<td>/ər/</td>
<td>96</td>
<td>110</td>
</tr>
<tr>
<td>/ər/</td>
<td>80</td>
<td>147</td>
</tr>
<tr>
<td>/əl/</td>
<td>136</td>
<td>68</td>
</tr>
<tr>
<td>/əl/</td>
<td>100</td>
<td>97</td>
</tr>
<tr>
<td>/əl/</td>
<td>100</td>
<td>120</td>
</tr>
<tr>
<td>/əl/</td>
<td>136</td>
<td>68</td>
</tr>
</tbody>
</table>

Dans les deux langues, lorsque /ə/ est en position initiale, le nombre d'espaces nécessaires pour reconnaître les voyelles /ər/ et /ər/ varie entre 350° et 660°, avec une moyenne autour de 500°. En position finale, cet intervalle varie entre 500° et 800° pour l'espagno (moyenne de 550°) et entre 500° et 800° pour le français.

Il faut toutefois noter que ces pourcentages élevés en français sont les fractions de durées relativement longues, puisque /r/ a de plus, selon un modèle de modélisation de la voyelle qui précède. Pour cette raison, les durées minimales nécessaires pour l'identification de /r/ donnent plus de voyelles que celles requises pour reconnaître /r/ isolé, à plus forte raison lorsque celui-ci est frappé, c'est-à-dire pour le segment /ər/ de 52 msec. En espagnol, cette différence ne s'applique pas en raison de la voyelle initiale /r/. En ce qui concerne la position intermédiaire, il faut noter la durée très longue du segment /əl/ chez un informateur canadien; la proportion d'espaces près de la limite du silence peut se traduire par une incertitude de /r/.

4. REMARQUES GENERALES

En vue de la poursuite de notre recherche, nous résumons les observations générales suivantes.

Une ligne ne peut jamais être identifiée en plan perceptuel sous la présence d'un noise pour une voyelle adjacente /ə/ est sensible dans les deux langues et se présente comme une modulation d'amplitude et de temps de la voyelle adjacente. /r/ est différent, puisqu'en espagnol, il se présente comme une interruption précédée d'un appui vocalique, alors qu'en français il produit des espaces durs, incluant ceux d'une fricative, bien que le plus fréquent que nous ayons observé soit la modulation de la voyelle adjacente analogue à celle produite par /l/.

La durée minimale nécessaire pour percevoir la voyelle est plus grande, lorsque celle-ci est précédée d'un silence de /ə/ que lorsque celles sont à l'initiale. Ceci s'observe dans les deux langues et dans les trois consoantes vocaliques de ce qui concerne /ə/, dans le cas de /ə/, avec différence moins accentuée dans les deux consoantes françaises, mais en espagnol, elle tend à se limiter au contexte de /ə/.

La conformation de /ə/ est différente des consoantes de /r/ et /ə/, lesquelles sont similaires entre eux. La voyelle /ə/ et la liquide restent à se différencier à l'identification minimale nécessaire pour l'identification de celle-ci. Il peut même arriver que cette identification ne soit possible qu'au niveau du mot.

En espagnol, /r/ et /r/ peuvent présenter des similitudes en position finale, à cause de l'absence d'interruption dans le /r/. Les deux liquides demeurent toujours bien différentes en position initiale. En français, les similitudes ou les différences peuvent présenter entre elles les deux liquides sont indépendantes des positions.

5. REFERENCES


76
PERCEPTION OF ANTICIPATORY VCV-COARTICULATION: EFFECTS OF VOWEL CONTEXT AND ACCENT DISTRIBUTION

Vincent J. van Beuven & Marc Ch. Dupuis
Dept. Linguistics/Phonetics Laboratory, Leyden University, The Netherlands

ABSTRACT
This paper examines the perceptual effects of first and second order coarticulation in VCV2-sequences in meaningful Dutch phrases, where the CV2-portion was deleted from the stimulus. V1 was /a,i,u/ or schwa and C was /p,t,k/. Either V1 was accented and V2 was not, or vice versa. Effects are generally stronger when V1 is unaccented. Identification of V2 but not for C is better from schwa than for other types of V1. The effects of accent distribution and vowel type are additive.

1. INTRODUCTION
By first order coarticulation we mean the mutual influence of adjacent phones. When a segment contains influence for a non-adja- cent phone we are dealing with higher order coarticulation. Generally, first order coarticulation is quite strong, and more easily demonstrated than higher order effects. Nevertheless, it has been shown that coarticulation effects can manifest themselves across several segment boundaries. Ohman [2] showed that part of the behaviour of the formant transition movements in V1 toward C in VCV2 sequences depends on the formant frequencies of V2 (and vice versa). Lip rounding in anticipation of a vowel can begin as many as four segments ahead (for a literature survey pertaining to these and subsequent claims cf. [3]). Additional evidence for the relatively large number of segments across which anticipatory coarticulation can extend is provided by investigations into anticipation of nasality.

Perceptual effects of coarticulation typically involve the use of stimuli of which parts have been deleted. The subjects' ability to identify the deleted sounds is considered a reliable measure of the perceptual usefulness of coarticulation. Stops turn out to be identified with above chance level on the basis of the transitions from, or into, the neighboring vowel. Similarly, it was demonstrated that consonants may contain perceptually useful cues for the identification of adjacent vowels. However, so far, no one has been able to show the perceptual relevance of higher order coarticulation effects using the truncation method. We claim that in none of the available studies assessing higher order coarticulation effects did the investigators include an optimal type of context for assessment of such effects. In the present experiment we set out to examine the perceptual effects of first and second order anticipatory coarticulation in VCV2 sequences under optimal conditions.

Vowels located in the central area of the traditional two-dimensional vowel diagram should be more prone to undergo the influence of context than vowels situated along the edges of such a diagram. Whereas the latter are accompanied by extreme tongue positions, the former are produced with the tongue in a more or less neutral position, from which it can move in any direction. We assume, therefore, that the central vowel schwa carries cues that are perceptually more useful than those carried by other vowels. We have tested perceptual effects of coarticulation in both schwa and the three point vowels. We predict higher identification scores for segments deleted after schwa than after /a,i,u/ (hypothesis 1).

We predict further that effects of coarticulation will depend on the distribution of stress over the coarticulatory domain. Stressed vowels may cause their features to spread further forward into following, and back into preceding segments than unstressed vowels. One therefore expects weak syllables to reflect coarticulatory influences from neighbouring stressed syllables more strongly than those which were not stimuli prepared from segments in which either V1 was accented and V2 unaccented, or V1 was unaccented and V2 was accented. Perceptual effects of anticipatory coarticulation will be stronger when V1 is weak and V2 strong, rather than vice versa (hypothesis 2).

Assuming additive effects of vowel quality and stress distribution, we further predict particular perceptual effects when V1 is both central and unaccented, and V2 is an accented point vowel (hypothesis 3).

2. METHOD
Targets were nine Dutch disyllabic words beginning with a CV1 syllable in which C was one of the three voiceless stops /p,t,k/ and V1 was one of the three phonologically long vowels /a,i,u/. The targets, such as tafel 'table' or koepel 'dome', were monomorphic words with lexical stress on their first syllable. Each target was embedded in a fixed set of carrier sentences, after one of four common, monosyllabic words. Since stress (to be realized as a pitch accent) was required either on the vowel of the monosyllabic word (V1) or the vowel of the target-initial syllable (V2), a total of 72 sentences (9 targets x 4 types of V1 x 2 stress patterns) was made.

The set of 72 sentences was read by a male native speaker of Standard Dutch. The final portions of the utterances were cut off in the silent interval of the voiceless pauses at the beginning of the target word. The resulting 72 sentences were copied on a test tape from nine series of eight sentences. In each series the order of the stimulus sentences was randomized. The interstimulus interval was fixed at 7s (onset to onset). Stimuli were presented through headphones to 62 native Dutch listeners. They were instructed to indicate which word they thought had been deleted after V1, with forced choice from nine preprinted response alternatives.

3. RESULTS
The experiment yielded a total of 62 (subjects x 72 stimuli) - 4,464 CV2 responses. The way in which consonant and V2 prediction was affected by the type of preceding vowel (V1) and the accent pattern over V1/V2 is shown in Table 1.

Table 1: Percent correctly identified C and V2 broken down by type of V1 and accent condition.

<table>
<thead>
<tr>
<th>RESPONSES FOR</th>
<th>V1</th>
<th>V2</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1 accented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V1= /a/</td>
<td>65</td>
<td>32</td>
</tr>
<tr>
<td>V1= /i/</td>
<td>62</td>
<td>38</td>
</tr>
<tr>
<td>V1= /u/</td>
<td>85</td>
<td>38</td>
</tr>
<tr>
<td>V1= schwa</td>
<td>80</td>
<td>41</td>
</tr>
<tr>
<td>V1 unaccented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V1= /a/</td>
<td>80</td>
<td>38</td>
</tr>
<tr>
<td>V1= /i/</td>
<td>64</td>
<td>32</td>
</tr>
<tr>
<td>V1= /u/</td>
<td>87</td>
<td>44</td>
</tr>
<tr>
<td>V1= schwa</td>
<td>82</td>
<td>50</td>
</tr>
<tr>
<td>Overall</td>
<td>76</td>
<td>39</td>
</tr>
</tbody>
</table>
C-identification

The overall correct identification score for C was 81% which is significantly above chance (33%). Obviously, the type of V1 played an important role in the identification of C. The vowels which were unaccented were, on the whole, identified best from preceding /a/. The overall effect of V1 on consonant identification was strongly significant (X²(3) = 185.5, p < .001). While subjects identified C significantly better from schwa (61% correct) than from /a/ (173% correct), the difference in scores between /a/ (86% correct) and schwa contexts was likewise found to be significant (X²(1) = 10.2, p < .01). Our first prediction, viz. that stops be better identified in the environment of preceding schwa than after point vowels, was therefore not quite confirmed by the overall results of V1-identification.

When we next examine the effect of accent pattern over V1/V2 it turns out that the results support our second prediction: with the accent on V2 rather than on V1 an overall score of 78% was found, when the stress distribution is reversed the overall score is 73% (X²(1) = 15.5, p < .001).

V2-identification

The vowels /a/ and especially /a/ were identified well above chance while identification of /i/ was not. The correct identification score is 39%, which is significantly different from chance (z = 12.3, p < .001; binomial test). Clearly, anticipatory coarticulation in word-final vowels (V1) can be usefully employed in the perception of non-adjacent vowels (V2).

The overall effect of V1 on the identification of V2 was substantial (X²(3) = 32.8, p < .001). Identification was significantly better when V1 was schwa (45% correct) than when V1 was /i,a,u/ (between 35 and 41% correct). This finding provides evidence that hypothesis (1), which predicts larger perceptual effects of anticipatory information in tokens of schwa than in tokens of point vowels, was correct (X²(1) = 5.9, p < .05). We conclude from hypothesis 2, where unaccented vowel tokens were expected to carry perceptually more relevant cues for the perception of V2 than were accented vowel tokens, is confirmed. Crucially, a large difference (41% versus 50% correct) between the two accentuation conditions can be observed in contexts where V1 was schwa (X²(1) = 8.3, p < .01). Our results suggest that identification of V2 was concerned, hypothesis (3), which predicts that facilitation of vowel identification should be maximal in the context of an unaccented schwa followed by an accented target- initial syllable, stands.

4. CONCLUSIONS AND DISCUSSION

We predicted larger percentages of correct identification for tokens of the central vowel schwa than from tokens of point vowels. The prediction was confirmed as regards identification of the deleted transconsonantal vowel; it could not be fully confirmed for the identification of the deleted consonant. Indeed, we have found that percent correct scores were of equal magnitude in the environment of preceding /schwa, a, i/ which were both 81% and 41% respectively. As concerns the role of V1 with respect to the identification of V2, our results demonstrate the expected effect: of the four vowels /i,a,u,schwa/ the central schwa most strongly facilitated the restoration of V2. Correct responses were generally more frequent in contexts where the vowel containing the anticipatory cues was unaccented and the target vowel was accented than in contexts where the accent distribution was reversed. This pattern of results was consistently found for first order (C) and second order (V2) coarticulation effects. Our experiment therefore provides substantial evidence that prediction (2) as stated in the introduction is essentially correct.

Moreover, our results indicate that the effects of stress distribution and V1 vowel type are largely additive. Crucially, vowel restoration was optimal when the target V2 was accented and when V1 was unaccented and schwa. Consequently, our hypothesis 3, predicts sensitivty of stress distribution and vowel type, stands. This experiment is first to show convincingly that perceptual effects of anticipatory coarticulation from-vowel onto-vowel are not necessarily restricted to immediately adjacent segments. Whenever conditions are carefully chosen, the perceptual effect of the second order vowel-on-vowel effect can be substantial. Clearly, why other researchers have by and large failed to uncover convincing perceptual effects of vowel-on-vowel coarticulation (3, 4, 6), lies in their infelicitous choice of stimulus material. Notice, in this context, that our optimal condition (predicting an accented V2 from a preceding unaccented schwa across an intervening word-initial stop consonant) is by far the most frequent triphone type in Dutch (and probably in English as well). This means that such coarticulation effects have ample opportunity to be utilized contrary to everyday speech perception.

NOTE

We thank S.G. Nooteboom and M.E.H. Schouten for comments. This research was partly supported by the Foundation for Linguistic Research, which is funded by the Netherlands Organisation for Research. NWO, under grant 350-161-023.

5. REFERENCES

EFFECT OF VOWEL QUALITY ON PITCH PERCEPTION

L. Raimo, O. Aalto, Å. Hellström* and E. Vilkman

Phonetics, University of Turku, Finland
Psychology, Stockholm University, Sweden

ABSTRACT

The intrinsic F0 (IF0) phenomenon was hypothesized to cause expectations of different pitches for different vowels. Listeners judged for pairs of synthetic vowels which had the higher pitch. The judgments were clearly based on vowel quality; there were also heavy effects of the time-order. The results can be explained by vowel-specific expected F0. This supports the view that intrinsic F0 of vowels is centrally controlled.

1. INTRODUCTION

Many explanations have been given for the intrinsic F0 (IF0) of vowels: under comparable circumstances, the high vowels [u, i] are produced with a higher F0 than the low vowels [æ, ã], cf. [7]. According to the acoustic coupling hypothesis, F0 is affected by vowel-specific changes in vocal tract acoustics. Mechanical coupling hypotheses suggest that IF0 depends on physiological interaction between the articulatory and the phonatory systems. From the results of our own acoustical and physiological experiments [8] we have concluded that none of these hypotheses is entirely satisfactory.

It has recently been suggested that IF0 is not merely a passive reflection of the biological characteristics of speech mechanisms, but centrally controlled [4]. This suggestion is supported by preserved IF0 in the esophageal speech of laryngectomized patients [7]. If IF0 is learned and automatized in language acquisition, then listeners may have different expectations for vowel pitches which in turn may cause pitch perception to depend on vowel quality. The present study tested this hypothesis experimentally.

2. PROCEDURE

The Finnish low vowels [a] and [æ], and the high vowels [u] and [i] were synthesized using the catharact type synthesis. All vowels had the same input amplitude configurations and the following formant structures (Hz) and method-dependent relative amplitudes:

<table>
<thead>
<tr>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>700</td>
<td>1100</td>
<td>2500</td>
</tr>
<tr>
<td>æ</td>
<td>650</td>
<td>1700</td>
<td>2500</td>
</tr>
<tr>
<td>u</td>
<td>300</td>
<td>600</td>
<td>2500</td>
</tr>
<tr>
<td>i</td>
<td>300</td>
<td>2250</td>
<td>2850</td>
</tr>
</tbody>
</table>

The durations of all vowels were 23 ms. Five F0 levels (1-5) were used. For Level 1, F0 was 104 Hz initially, reaching 114 Hz after 9 ms and then declining to 84 Hz. Levels 2, 3, 4, and 5 deviated at all points from level 1 by +3, +6, +9, and +12 Hz. Levels 2 and 4 were used as the first members and all five levels as the second members of pairs. Thus the largest differences within the pairs were 9 Hz (more than a semitone). 1.1 s intervened between the members of each pair and 3.6 s between the pairs. All possible vowel pairs, 168 vowel pairs in all, were recorded in random order and presented to listeners who had to judge for each pair which vowel had the higher pitch or if they had equal pitch. For each vowel combination, 20% of the pairs had equal F0. In 40% the first vowel was higher, in 40% the second. There were two groups of Finnish-speaking listeners: 32 university language students (4 men and 28 women, mean age 22) and 66 members (29 men, 37 women, mean age 38) of a well regarded amateur symphonic choir, who were thought to be more than normally trained in discriminating vowel pitch.

3. RESULTS

In terms of correct judgments, the choir performed slightly better than the students: For the eight pairs in which equal-quality vowels were juxtaposed with the maximal (9 Hz) F0 differences, the choir made 64% and the students 31% correct judgments. Below, the percentages of "equal" judgments are given for the vowels in each combination (mean over the two time orders). The percentages of "equal" judgments are given in parentheses. (For the students, each row represents 640, for the choir, 1320 judgments.)

Students Choir
æ - u 79-9 (12) 64-24 (13)
a - ã 75-8 (17) 56-27 (17)
i - ë 68-9 (23) 56-27 (18)
æ - i 49-24 (27) 45-37 (18)
a - a 40-31 (29) 53-28 (19)
a - i 43-35 (22) 39-42 (19)

Thus, in both groups, [æ] was heard as higher and [u] as lower compared with any other vowel.

For all vowel combinations, the groups made the time-order dependent judgments:

Students Choir
V1-V2 28-41 (31) 30-45 (25)

Thus, the second vowel was heard as the higher more often than the first. This is called (see [2]) a negative time-order error (TOE). There were, however, clear differences between the vowels, as well as between the groups:

Students Choir
æ - æ 12-35 (54) 17-37 (46)
a - a 11-34 (55) 17-33 (50)
i - ë 23-18 (59) 17-34 (69)
u - u 25-8 (66) 23-26 (50)

Thus, for both groups, the higher in pitch a vowel was judged when compared with other vowels, the stronger was its tendency to be judged as higher when second in a pair and compared with itself. For [æ] with the students, the negative TOE was reversed to positive.

For describing and explaining TOEs (which are found for many kinds of stimuli including tonal...
loudness and pitch), Hellström [2] developed a general model for stimulus comparisons. According to this model, the two pitches are not compared directly; their mean judged difference (as measured, e.g. by D%), the difference between the percentages of "first higher" and "second higher" judgments) is proportional to the difference between two compounds. In the present case, each compound corresponds to one of the vowels in the pair, and is a weighted sum of its actual pitch, with relative weight s, and its expected pitch (its adaptation level, AL), with relative weight t-s (NB: s may be either >1 or <1).

Assuming identical, linear relations between F0 and pitch for all vowels in the small F0 range used, the expected F0 (AL) (in Hz relative to the mean F0, 106 Hz) and s values (up to a scale constant, k) could be estimated by multiple linear regression of D% for each pair on its F0 values (R was .954 for the students, .892 for the choir):

Students

<table>
<thead>
<tr>
<th>Vowel</th>
<th>k1</th>
<th>k2</th>
<th>AL</th>
<th>k1</th>
<th>k2</th>
<th>AL</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4.0</td>
<td>7.1</td>
<td>-16</td>
<td>4.0</td>
<td>7.1</td>
<td>-13</td>
</tr>
<tr>
<td>ä</td>
<td>3.3</td>
<td>5.2</td>
<td>-4</td>
<td>3.6</td>
<td>4.9</td>
<td>-13</td>
</tr>
<tr>
<td>i</td>
<td>0.8</td>
<td>3.6</td>
<td>-7</td>
<td>4.1</td>
<td>5.4</td>
<td>-2</td>
</tr>
<tr>
<td>u</td>
<td>3.1</td>
<td>5.0</td>
<td>-14</td>
<td>3.7</td>
<td>5.3</td>
<td>-9</td>
</tr>
</tbody>
</table>

For both groups and all vowels, the vowel's weight when first in a pair (s1) was higher than its weight when second (s2). For the students, AL (expected F0) was highest for [u] and lowest for [ä]. For the choir, [u] had a higher AL than the other vowels.

4. DISCUSSION

The purpose of our study was to test if perception of vowel pitch depends on vowel quality, because in articulation pitch varies with quality. The result was clearly positive: other things being equal, the vowels [ä] and [a], which have low IF0s, were heard as highest in pitch. The results cannot be explained by the amplitude differences, as we found no clear relation between amplitude and experienced or expected pitch. The distribution of energy in the vowel spectra might be of greater importance.

However, our results indicate that the most important factor for vowel-specific pitch is expected F0, which is higher for the high than the low vowels. By reference to Hellström's [2; 3] model, the different AL and 8 values explain both all vowel-specific pitch differences and the TOEs in our data. Besides, pitch discriminability (indicated by k8) in the student group was much poorer for [u] than for the other vowels. The results thus clearly support our hypothesis that because in articulation F0 varies between vowels, perceived vowel pitch depends on vowel quality.

It is interesting to note that in another recent study [9] the vowel [u] was produced with higher subglottal pressure than the other vowels [i, ä]. Thus the vowel [u] seemed to be different from other vowels also in terms of the physiology of speech production. These effects may have been emphasized by the especially dark quality (low F2) of Finnish [u]. The question whether our findings share a common basis, e.g. higher respiratory effort in production perceived as stress, remains open for speculation.

Our study also supports the view that IF0 is an inherent property of vowel prototypes in the brain; even trained singers cannot eliminate its effect on the perceived pitch. In vowel pitch perception, then, the IF0 behaves somewhat like formants, which are not perceived separately, but as integrated characteristics of vowel quality. IF0, nevertheless, has no phonologically distinctive function in languages [5]. Our results are in accordance with those speech perception theories which maintain that speech perception is based on articulatory rather than acoustic parameters of speech sounds; see [6].

5. ACKNOWLEDGMENTS

We express our deepest gratitude to the Speech Transmission Laboratories of the Royal Institute of Technology in Stockholm and especially to Rolf Carlson and Björn Granström for their invaluable help in producing the synthetic vowels.

6. REFERENCES


THE PERCEPTION OF SILENT-CENTER SYLLABLES IN NOISE

Susan L. Hura

University of Texas, Austin, Texas

ABSTRACT

The perception of vowel-less /b/-vowel-/v/ syllables was tested at various signal to noise ratios. Contrary to what has been shown in previous studies [7], vowels in these "silent-center" syllables were not identified at the same accuracy as vowels in full syllables. This calls into question the degree to which the perception of silent-center syllables can be seen as evidence for the theory of dynamic specification.

1. INTRODUCTION

Dynamic specification is a recent theory of vowel perception proposed by Strange [7] in which vowels are conceived of as gestures having intrinsic timing parameters. Dynamic specification is in opposition to a traditional target theory which states that vowel recognition is based upon characteristic frequency values for the first 2 (or 3) formants taken from a single time slice in the syllable nucleus. Strange cites certain perceptual results in support of dynamic specification. Of specific interest here is the correct identification of vowels in vowel-less syllables. Using a waveform editing technique, "silent-center" syllables were generated, in which the vowel nucleus was attenuated to silence, leaving 3 or 4 pitch periods of consonant transition at either edge of the syllable. Strange found that subjects were able to identify the vowels in silent-center [SC] syllables with nearly the same accuracy as they could the vowels in full syllables, thus refuting a simplification theory. If recognition can proceed in the absence of vowel nucleus information, then this information is not the determining property of vowel identity.

In this paper, Strange's SC result is reconsidered. Let us begin with the assumption that target formant values, formant transitions and other dynamic attributes all play a role in the identification of vowels. These factors normally provide redundant and overlapping information about vowel identity, thus it is not surprising that identification can be relatively accurate in the absence of some of this information. SC syllables are an example of a stimulus where some vowel information is absent, but a great deal remains. In favorable listening situations, it is possible to make up for the lack of one sort of information by focusing on remaining information. Because vowel identification is a familiar task and an experimental setting is relatively free of distractions and ambient noise, the listening conditions in Strange's experiment were close to ideal. Under degraded listening conditions, however, listeners may rely more on each source of information than they otherwise would. My claim, then, is that Strange's SC result is due to the favorable listening conditions under which she tested identification performance. To investigate this claim, Strange's Experiment 3 [7] was partially replicated.

2. PERCEPTUAL STUDY

2.1. Stimuli

Stimulus materials consisted of /b/-vowel-/v/ syllables in the carrier phrase "I say the word /b/ vowel/some more," for each of 10 vowels. The speaker was an adult male with a midwestern dialect. Stimuli were digitized and waveform edited to produce SC syllables according to criteria defined by Strange [7]. Full and SC syllables were then embedded in wide-band noise. Two sorts of SC syllables in noise were created: in the first (SC1), the amplitude of the initial and final components has been boosted such that their peak amplitude is equal to the full syllable peak; in the second set (SC2), amplitude of initial and final components is the same as for full and SC versions of a syllable at the same S/N. Six S/N were created for each syllable type for each /b/ by varying the amplitude of the signal in relation to constant amplitude noise. All stimuli were embedded into the same carrier phrase.

2.2. Subjects & Procedure

Stimuli were randomized, and a listening test was created. Stimuli were presented in blocks of 21; interstimulus interval was 4 seconds, interval between blocks was 8 seconds. The task of the subject was to circle the /b/ word they heard on a preprinted answer sheet. There were a total of 252 items in the test—labeled approximately 30 minutes. The 26 subjects tested were U.T. undergraduate volunteers who were paid for their participation. 20 of the subjects spoke a Texas dialect; dialects of the remaining 6 varied. Subjects were tested in a laboratory setting in groups of 3 to 6.

2.3 Results & Discussion

Table 1 below gives percent correct by syllable type and S/N, collapsed across vowels. Examining this data it is apparent that vowels were perceived more accurately in full syllables than in either SC1 or SC2 syllables. Figure 1 shows these same results graphically.

A two-way analysis of variance on syllable type and S/N was performed; both of these factors were shown to have an effect, but their interaction does not. Syllable type (full versus SC1 versus SC2) is significant for F (2, 162) = 22 at p < .025; S/N is also significant for F (5, 162) = .81 at p < .001. T-tests for differences among the means of the 3 conditions were performed, which showed that full syllables are significantly different from either type of SC syllable (p < .025), and that the two types of SC syllables are not significantly different. Thus it appears that given a more difficult identification task, vowels are significantly more difficult to perceive in SC syllables.

Table 1: Overall percent correct by syllable type.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>3</th>
<th>6</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>60.3</td>
<td>74.2</td>
<td>85.0</td>
<td>87.7</td>
</tr>
<tr>
<td>SC1</td>
<td>53.1</td>
<td>48.8</td>
<td>76.5</td>
<td>86.3</td>
</tr>
<tr>
<td>SC2</td>
<td>56.5</td>
<td>58.1</td>
<td>66.4</td>
<td>73.5</td>
</tr>
</tbody>
</table>

Figure 1: Overall percent correct by syllable type.
3. CONCLUSIONS

The results of this study show that while it is easy to identify the vowels in silent-center syllables, this identification is not as accurate as for full syllables. It was shown that under degraded listening conditions, when the listener is more dependent upon the redundancies of the speech signal, identification performance is significantly better for full syllables than for silent-centers. The ability to accurately perceive a vowel in a syllable where it one is not physically present is certainly remarkable. The current data show that even at low S/N, subjects identify vowels in SC syllables at well above chance level. However, given the poorer identification performance on SC syllables in difficult listening situations, we are not justified in claiming that transition information has greater importance than the nucleus in the specification of vowels. This is not to say that nucleus information is privileged, for the representation of a vowel as a static point in F1/F2 space is clearly insufficient to account for the present results. It is clear, however, that syllables containing nucleus information are better perceived than those without it.

Perhaps a dual-target model of vowel specification [1] can provide an explanation for the current results. If a vowel is specified by formant values in both the nucleus and offglide, identification should be more accurate when both of these are present in the stimulus, as is the case for full syllables. When some of this information is missing, as in SC syllables, a dual-target model predicts the poorer identification performance shown here.

4. REFERENCES


MINIMAL DURATION FOR PERCEPTION OF FULL-SPECTRUM VOWELS

Rudolf Weiss  W. S. Brown, Jr.  Shaw N. Gynan
Western Washington University  University of Florida  Western Washington University
Bellingham, WA, USA  Gainesville, FL, USA  Bellingham, WA, USA

ABSTRACT
This study developed a perception test to determine the minimal duration threshold levels of vowels on the basis of short bursts of complete waveforms (full spectral cues) of five vowels [i e a o u].

1. INTRODUCTION
Although considerable and meaningful work has been done in the area of vowel perception over the last several decades, recently developed and fairly accessible instrumentation is now available which allows for relatively easy access and manipulation of the speech signal [3]. Different approaches have been used such as short bursts of vowels at set-time intervals with manipulation of F₁ and F₂ frequencies [2,3]. Other studies have involved masking techniques [5] and still others have dealt with vowel formant transitions for vowel vs. consonant in semi-vowel identification [4]. Most vowel perception experiments share in common the fact that they use synthesized vowels with manipulations of F₁, F₂ and/or F₃ relative to each other in frequency, bandwidth and/or synchrony. Shortcomings of some of these models have been shown by Bladon [1]. Since hitherto most experiments have dealt with synthesized vowels and manipulations of the spectra in efforts to isolate specific functions of distinct acoustic cues, it was decided to experiment with complete waveforms (full spectral cues) of steady-state portions of vowels to determine on the basis of short bursts the minimal durational thresholds for consistent vowel classification. It was hoped that acoustic cues, it was decided to experiment with complete waveforms (full spectral cues) of steady-state portions of vowels to determine on the basis of short bursts the minimal durational thresholds for consistent vowel classification. It was hoped that we could also thereby ascertain something about the degree of difficulty in vowel perception as the time duration of bursts decreased, i.e., to verify through other means that high vowels [i] and [u] are generally easier to classify as maintained in Lieberman [5] and as found in previous cross-language studies by Weiss [7,8], showing that durational variation affects the high vowel [i] less than other vowels.

2. PROCEDURE
Five vowels [i e a o u] were produced in steady-state fashion by a male speaker (F₀ = 100 Hz ± 2 Hz) and a female speaker (F₀ = 201 Hz ± 3 Hz). These vowels were digitized using the MacSpeech Lab II/MacAudio II hardware/software program. A sampling rate of 44KHz was used in the recording of the utterances which yielded a frequency response ceiling of 20 KHz. Using built-in routines of the MacSpeech Lab program, the utterances were equalized in amplitude and segmented on the basis of full-wave displays. They were then segmented first into 300 ms segments (which served as the reference cue in the perception tests) and then into smaller whole-wave units. The formant distribution figures (LPC) for both the male and the female utterances are given below:

Male: F₀ F₁ F₂
[i] 100 285 2405
[e] 101-102 408 2242
[a] 98-99 652 1019
[o] 101-102 489 775
[u] 99-101 285 775

Female: F₀ F₁ F₂
[i] 201-204 285 2691
[e] 198-201 449 2405
[a] 201-203 530 1223
[o] 199-200 245 571
[u] 201-203 408 775

Segments were cut from the mid-point of each vowel. From the male speaker sample segments of increments from one to four complete cycles yielded four samples in duration from 10 to 40 ms. A parallel procedure was followed for the female speaker. However, since the F₀ was twice that of the male, one to eight complete cycles yielded samples in duration from 5 to 40 ms. In addition, a one-half cycle segment of each vowel beginning with the first positive rise of the wave was isolated, yielding additional segments of 5 ms for the male and 2 ms for the female. Thus the male voice yielded five segments of each vowel for a total of 25 segments. Two tests were developed: one for each voice, in which each token occurred three times. This resulted in two perceptual test tapes: one of 75 tokens for the male voice and one of 135 tokens for the female voice. The tokens were randomized and rerecorded at five-second intervals to minimize the effect of short auditory memory. For reference purposes, two repetitions of 300 ms tokens of each vowel for the male and female voice were given at the onset of each test.

Both tests were administered individually to 38 phonetically unsophisticated subjects, 16 males and 22 females, at the University of Florida. The mean age of the subjects was 20. The order of presentation of the two tests was reversed for half of the subjects.

3. EQUIPMENT
Digitizing was performed with a Mac II with 4 mb RAM and a 68020 microprocessor with a Mac Speech Lab II/MacAudio II hardware-software package. Samples from the digitized utterances were made with a Teac V-570 cassette deck. The listening tests were administered individually using a Teac W370C cassette deck in conjunction with a Technics SU-V450 integrated amplifier and a Technics Model SC-C36 two-way speaker system for the reference samples.

4. RESULTS
The results indicated a high degree of accuracy in perception of vowels of most durations. Variations in responses to individual vowels were significant only for the shortest durations. Even a one full-spectral wave cue (female - 5 ms/male - 10 ms) was long enough for fairly consistent classification. The lengthy interval of 5 ms between cues no doubt enhanced categorical perception by minimizing short auditory memory as predicted by Repp [5]. There was still sufficient cue information even if only half the spectral information for one wave form was given to enable fairly consistent identification of vowels.
It is questionable how meaningful a ranking order of vowel difficulty might be due to the high degree of correct classification of responses. However, based on a possible 1026 correct classifications of each female vowel and 570 possible correct classifications of each male vowel, the ranking order from easiest to most difficult vowel for each voice is indicated below. Percentage indicates the total errors made by all subjects to each vowel.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Male (%)</th>
<th>Female (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[o]</td>
<td>2.1%</td>
<td>7.6%</td>
</tr>
<tr>
<td>[u]</td>
<td>5.6%</td>
<td>8.0%</td>
</tr>
<tr>
<td>[i]</td>
<td>5.8%</td>
<td>9.7%</td>
</tr>
<tr>
<td>[e]</td>
<td>5.9%</td>
<td>24.0%</td>
</tr>
<tr>
<td>[æ]</td>
<td>14.9%</td>
<td>35.5%</td>
</tr>
</tbody>
</table>

It is obvious from the above statistics that the most difficult male vowel to categorize was [æ], with 14.9% errors, and the most difficult female vowel to categorize was [u] with 35.5% errors. Thus prior findings that [i] and [u] are among the easiest vowels to classify are not supported by this study. It is also apparent that the female vowels posed much greater perceptual difficulties even if only vowels of the same duration are compared. The table below illustrates comparable male/female token values. For each time variation there were 114 tokens for 38 subjects.

<table>
<thead>
<tr>
<th>Duration</th>
<th>Male Error Percentage</th>
<th>Female Error Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 ms</td>
<td>74.6% (49.2-84.3%)</td>
<td>65.4% (29.3-90.5%)</td>
</tr>
<tr>
<td>5 ms</td>
<td>74.7% (48.3-90.4%)</td>
<td>68.5% (25.1-90.6%)</td>
</tr>
<tr>
<td>10 ms</td>
<td>79.7% (40.4-93.0%)</td>
<td>78.7% (23.7-90.3%)</td>
</tr>
<tr>
<td>15 ms</td>
<td>84.8% (57.9-98.2%)</td>
<td>87.4% (65.0-99.2%)</td>
</tr>
<tr>
<td>20 ms</td>
<td>87.4% (65.0-99.2%)</td>
<td>92.8% (84.3-99.2%)</td>
</tr>
<tr>
<td>25 ms</td>
<td>92.8% (84.3-99.2%)</td>
<td>95.7% (76.7-99.5%)</td>
</tr>
</tbody>
</table>

For context independent recognition of vowels the male voice obviously yields the best response. With the exception of [æ] all vowels could be truncated to one wave form (20 ms) and still have 90-100% recognition. For the female voice even 2 wave forms (10 ms) would yield only 40% recognition for [u] but 85-93% for all other vowels.

This study shows that overall best results for vowel recognition occurs for two wave shapes (20 ms) for the male voice with recognition level of 95.7% (minimum of 90.4% for any vowel); for the female voice the best results are with five wave shapes (25 ms) with a recognition level of 92.8% (minimum of 84.3% for any vowel). Thus it appears that duration, not number of complete cycles, is an overriding factor in determining minimal threshold levels in perception. The threshold for highly accurate classification seems to be located at between 20-25 ms.

Analysis of variance failed to establish significant correlations regarding vowel formant spread or the effect of order of presentation. Nor could statistically significant differences between male and female subjects in accuracy of vowel identification be established. A larger data base would be necessary to confirm this finding.

5. CONCLUSION
The degree of persistence of full-spectrum cues through the shortened time window was unexpected. A high degree of accuracy in vowel perception remained even to the shortest burst which allowed perceptual/auditory access only to half of a wave shape, i.e., a time duration of little more than 2 ms. Optimum results were obtained in the 20-25 ms token range. The implication of these preliminary findings is that if full-spectral cues are given, an exceedingly small time frame will suffice for fairly consistent and reliable perception and classification of vowels. More than twice as many errors were made in classifying the female tokens which correlated closely to the increase of the fundamental frequency of the female voice. We plan to expand our study to allow for a larger data base in forthcoming endeavors.

N.B. This research was made possible through the use of the research facilities at IASC, University of Florida.

7. REFERENCES

Table 1: Perception Errors of Comparable M/F Values

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Male Error Percentage</th>
<th>Female Error Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>[i]</td>
<td>21.9%</td>
<td>37.7%</td>
</tr>
<tr>
<td>[e]</td>
<td>9.6%</td>
<td>0%</td>
</tr>
<tr>
<td>[o]</td>
<td>10.5%</td>
<td>70%</td>
</tr>
<tr>
<td>[u]</td>
<td>51.7%</td>
<td>51.7%</td>
</tr>
</tbody>
</table>

The study shows that in general errors in perception increase as the vowel duration decreases. An exception is
PERCEPTION OF THE HIGH VOWEL CONTINUUM: A CROSSLANGUAGE STUDY

Bernard L. Rochet

University of Alberta, Edmonton, Alberta, Canada

ABSTRACT

An imitation task in which speakers of English and of Brazilian Portuguese repeated a randomized list of monosyllables recorded by a native speaker of Standard French, and containing the vowels /I, /y, /u/, and /o/ in different consonantal contexts. In the perceptual task, English and Portuguese subjects were asked to identify as /I/ or /y/ 3 sets of randomized synthetic stimuli containing a high vowel continuum (from /I/ to /y/), with each set consisting of isolated vowels and the other 2 sets of vowels in the environments /I_/ and /y_/ respectively. Only the results for the isolated vowel stimuli will be reported here. The French subjects which took part in the experiment were asked to identify each of the synthetic stimuli as one of the three vowels /I, /y, or /u/. The stimuli were synthesized in cascade at a 10-kHz sampling rate using X1's [5] cascade/parallel speech synthesizer. The vowel portion of the stimulus was 200 ms long and varied along the F2 dimension between 500 and 2500 Hz. The stimuli were produced by the fact that these speakers perceive and divide the high vowel continuum in different ways.

1. INTRODUCTION

A phenomenon familiar to second-language (L2) instructors is the inability of some learners to produce sounds of the target language not present in their native--or first (L1)--language inventory. For example, evidence from speech production reveals that, in attempting to speak a second-language whose inventory contains the French vowel /I/ and /y/, native speakers of languages whose inventory contains only the 2 high vowels /I/ and /y/ find it difficult at first to produce the target vowel /I/. When anything at all is done in the L2 classroom to correct this situation, the problem is usually addressed by means of articulatory instruction, and the students are advised to produce a high vowel which is at the same time front and rounded. The fact that, in spite of such straightforward instruction, beginners often go on mispronouncing the target vowel /I/, show a low rate of success in imitation tasks, and fail to detect any difference between their faulty pronunciations and the target sound, suggests that a faulty production of the target sound may be attributable--at least in part--to its faulty perception. This interpretation is not new, and it is inferred from production evidence in general, and imitation experiments in particular, that a sound occurring in L2 but not in L1 is judged to belong to an L1 category, a process labelled "interlingual identification" [3]. The purpose of this paper is to demonstrate that accented pronunciations of the French vowel /I/ by speakers whose native languages contain only the 2 high vowels /I/ and /y/ reflect the way such speakers perceive and divide the high vowel continuum.

2. PROCEDURE

This hypothesis was tested by means of an experiment consisting of an imitation task (to establish in a systematic way how each subject pronounced the target vowel /I/), and of a perceptual task (to establish how subjects divided the high vowel continuum in terms of the categories of their respective native languages). In addition to native speakers of Standard French, 2 groups of 10 speakers each (ranging in age from 23 to 32) took part in the experiment: speakers of Canadian English, who have been observed to replace French /I/ with an /u/-like vowel [9]; and speakers of Brazilian Portuguese, who have been observed to replace French /y/ with an /I/-like vowel.

In the imitation task, each subject was asked to repeat a randomized list of monosyllables recorded by a male native speaker of Standard French, and containing the vowels /I, /y, /u/, and /o/ in different consonantal contexts. In the perceptual task, English and Portuguese subjects were asked to identify as /I/ or /y/ 3 sets of randomized synthetic stimuli containing a high vowel continuum (from /I/ to /y/), with each set consisting of isolated vowels and the other 2 sets of vowels in the environments /I_/ and /y_/ respectively. Only the results for the isolated vowel stimuli will be reported here. The French subjects which took part in the experiment were asked to identify each of the synthetic stimuli as one of the three vowels /I, /y, or /u/. The stimuli were synthesized in cascade at a 10-kHz sampling rate using X1's [5] cascade/parallel speech synthesizer. The vowel portion of the stimulus was 200 ms long and varied along the F2 dimension between 500 and 2500 Hz. The stimuli were produced by the fact that these speakers perceive and divide the high vowel continuum in different ways.

3. ANALYSIS

3.1. Production (Imitation Task)

The items recorded for each subject during the imitation task were digitized and presented in randomized order to 3 native speakers of Standard French for evaluation on a 5-point scale: 1 = /I/ or /u/ vowel between /I/ and /y/; but closer to /u/; 3 = vowel between /I/ and /y/; but closer to /I/; 4 = /I/ or /u/-like vowel between /I/ and /u/; but closer to /y/; 6 = vowel between /I/ and /u/; but closer to /I/; 7 = /I/ or /u/-like vowel. On the basis of this scale, a score between 1 and 4 indicates a vowel between /I/ and /u/; and a score between 4 and 7 a vowel between /y/ and /I/. The stimuli were presented on-line on a Zenith 286 microcomputer, by means of sound cards at the University of Alberta, and delivered binaurally through TD-149 earphones. When they were not successful in repeating French /I/ as /y/ or an /u/-like vowel, Portuguese speakers repeated it 95% of the time as /I/ or an /u/-like vowel (generally a lax variant thereof), or as a vowel described by the subjects as falling between /I/ and /y/. They repeated French /y/ as /u/, an /u/-like vowel, or even a vowel between /y/ and /I/ only 5% of the time. Their mean score for these non-/y/ productions was 2.13. On the other hand, when English speakers did not succeed in repeating French /I/ as /y/ or an /u/-like vowel, they were found to repeat it as /I/ or an /u/-like vowel (a lax variant thereof), or as a vowel between /I/ and /u/ 92% of the time, and as an /I/-like vowel or a vowel between /I/ and /y/ 8% of the time. Their mean score for non-/I/ productions was 5.01. These results support observations that Portuguese speakers generally replace French /I/ with an /I/-like vowel, and that English speakers generally replace it with an /u/-like vowel [8].

3.2. Perceptual Task

The results of the perceptual task (both pooled and individual) were analyzed to yield crossover boundary values between adjacent vowel categories, and to produce graphs of the identification functions. As shown in Figs. 1 and 2, the crossover boundary between /I/ and /y/ is located much higher on the F2 scale for English speakers (1900 Hz) than for Portuguese speakers (1575 Hz). A comparison of the English and Portuguese labeling functions with those obtained from native speakers of Standard French (Fig. 3) shows that stimuli with F2 values between 1500 and 2100 Hz, which are identified as /I/ by French speakers, are most of the time labeled as /y/ by English speakers and as /y/ by Portuguese speakers.
for an explanation of the phenomenon of interlingual identification. When called upon to imitate French /ι/ L2 learners do not have access to French categorization functions, but to natural tokens of that vowel pronounced by native French speakers. To understand the process of interlingual identification, one must therefore relate

mean F2 values of the vowel /ι/ obtained from production data to the L2 learners’ identifications functions. The average value of F2 for French /ι/ has been given as 1850/900 Hz at the high end of the range [1] (2), and as 1675 Hz at the lower end [6]; the average F2 value of the French tokens is 1760 Hz, with extreme values of 1612 Hz and 1824 Hz. It can be seen from Figs. 1 and 2 that most tokens with such values fall within the bounds of the /ι/ category for Portuguese speakers, and within the /u/ category for English speakers.

4. DISCUSSION

4.1. The parallelism between the results of the imitaiton task and those of the perceptual task appear to support the hypothesis that accented pronunciations of L2 sounds by untrained learners may be perceptually motivated. It suggests that, in early stages of L2 learning, learners perceive L2 sounds in terms of their native phonetic systems, which may be perceptually motivated. Thus, the categorization of L2 phonemes as acoustically different from those of their native language occurs in response to the perceptual boundaries of the same L1 category, even if the tokens of the native phonetic system are acoustically the same or different. Once assigned to that category, the perceptual boundary must exist between these two vowels.

1. The identification functions represented in Fig. 2 indicate that, although English has only two high vowel categories labelled /u/ and /u/, the perceptual boundary between these two categories typically coincides with the perceptual category between /u/ and /u/ in French (see Fig. 3) and in Swedish [8]. It seems likely, therefore, that the discrimination peak observed by Stevens et al. for their English subjects occurred not because of a natural perceptual boundary in the real world, but because the stimuli were perceptually bounded into two separate categories. In addition, a comparison of the English and Portuguese identification functions (see Figs. 1 and 2) shows that the perceptual boundary between the high vowels /u/ and /u/ does not occur in the same location in different languages, and suggests that the location of this perceptual boundary in languages having only two high vowels is the result of linguistic experience rather than a reflect of some basic property of the auditory mechanism.

5. REFERENCES

UNDERSTANDING DISFLUENT SPEECH: IS THERE AN EDITING SIGNAL?

Dept. of Linguistics and Centre for Speech Technology Research, University of Edinburgh, Scotland

ABSTRACT

The problems posed by the frequent occurrence of disfluency in normal speech are important both for psycholinguistic and computational models of speech understanding. The most basic of these problems is determining when disfluency has occurred. Hindle (1) makes use of a phonetic 'editing signal' which marks the end of the material to be ignored and indicates the onset of the repair. This paper presents the results of gating experiments on spontaneous speech which show that only a minority of disfluencies can be detected by the point where this signal is clamped to occur, but that nearly all are obvious to listeners within the first word of the repair.

1. INTRODUCTION

Unlike written or read language, spontaneous speech is characterised by numerous disfluencies. For the purposes if this discussion, disfluency will be understood to consist of two main types: repetitions (Example 1) and false starts (Example 2).

Both may be of lengths varying from less than a syllable to several words. Other hesitation phenomena - silent and filled pauses and lexical fillers - will not be discussed.

Example 1: Repetition:

'And you'd re read really need about eight...'

Example 2: False Start:

'Because although the hell the rules say that...'

It is all too easy to miss disfluencies when transcribing spontaneous speech verbatim, and all too difficult to believe that so many occurred when perusing a correct transcription because we appear to notice very few of them as they occur.

One of the factors which may facilitate the processing of disfluent speech could be the presence of cues in the speech stream prior to the break in fluency which prepare listeners for a break. Don Hindle (1) makes use of this idea in his algorithm for parsing speech with disfluencies:

"Two features are essential to the self-correction system: 1) a very self-correction site [...] is marked by a phonetically identifiable signal placed at the right edge of the expungation site..."

(1) p. 128

Hindle's editing system depends crucially on the presence of this editing signal (see Labov (2)), defined as [1]. The system takes as input a transcription in standard orthography of conversational speech which has editing signals inserted by the transcriber, when noted, at the point of interruption.

The experiments described in this paper are designed to establish the location of the editing signal to a first approximation. They use materials from a sample of repetitions and false starts drawn from and representative of those in a corpus of studio-recorded spontaneous conversational English. The first experiment establishes that listeners are able to recognise that an utterance is disfluent by the offset of the first word following a disfluent interruption. The second experiment addresses Hindle's supposition that an editing signal 'placed at the right edge of the expungation site' (i.e. immediately following the section of speech that is to be ignored and prior to the onset of the continuation) indicates to the listener that a disfluency is present. It is found that the majority of disfluencies are not detectable at this point in the utterance. The conclusion is reached that, if an editing signal is present in disfluent speech, it is not as a discrete phonetic signal, but rather a feature of the prosodic disruption that takes place.

2. EXPERIMENT ONE

2.1. Introduction

This experiment was designed to test the hypothesis that disfluency can be recognised by the offset of the word following the interruption point.

2.2. Materials

From a corpus of spontaneous speech, recorded digitally in a studio, 30 spontaneous disfluent utterances were selected, each containing a token of one of a set of types of disfluency, to be used as test items. The types of disfluency and the numbers of each type used were representative of the distribution of types of disfluency identified in the corpus by the first author. Test items were divided equally among the six speakers whose conversations make up the corpus.

Next, another 30 utterances were chosen from the corpus to provide spontaneous fluent controls for the disfluent items. These items were selected to match the disfluent utterances for structure, length and prosody as far as possible.

To provide controls better matched in structure to the spontaneous disfluent utterances, each such item was edited using ILS to remove the disfluency and leave, without interruption, the fluent parts of the utterance. Each of the original speakers then heard the doctored versions of his or her utterances and was asked to produce 6 fluent imitations of each. The speakers' responses were recorded under the same conditions as in the recording of the original conversations. For each item, the most accurate of the imitations was selected to be the control for that item, accuracy being defined as closest matching in terms of rate and rhythm of production.

Examples of the resulting test materials are given below.

Example 1:

Spontaneous Disfluent:

'... it's quite obvious he's he's on something...'

Rehearsed 'Disfluent':

'... it's quite obvious he's on something...'

Spontaneous and Rehearsed Fluent:

'... we know that it's not going so...'

All the utterances to be used were sampled on ILS on MASSCOMP through a 8kHz filter at 20kHz, together with up to 10 seconds of the conversation which occurred prior to the test utterance, which provided some discourse orientation. The onset of each word in each item was determined from a combination of auditory information and time-amplitude waveform. Each item was then gated at word boundaries so that the first stimulus for an item ran from its onset to the end of its first word (its'), the second from its onset to the end of its second word (it's quite), the third to the end of its third word (it's quite obvious) and so on.

The test materials were divided into two complementary sets of sixty utterances so that neither of the two sets of subjects heard both the spontaneous and the rehearsed versions of any utterance. Each set of 60 items was blocked by speaker and recorded on a separate test tape.

2.3. Subjects and Procedure

Twenty students and staff members of the University of Edinburgh served as subjects, 10 per group. All were native speakers of English familiar with the range of accents represented in the
experimental materials and all reported having normal hearing.

The experiment was run in two sessions of approximately 45 minutes.

Subjects were given adequate time to familiarise themselves with each speaker's voice and all utterances were presented with about ten seconds of the dialogue prior to the utterance.

There were two tasks in the experiment: word recognition and disfluency recognition. For the word recognition task, subjects were asked to write down after each gazed presentation what they thought the latest word presented was and to make any amendments required to previous words in the appropriate part of the answer sheet. For the disfluency recognition task, subjects were asked to make a judgement on a 1-5 scale about whether they considered that the utterance was fluent at the current word gate. A score of 1 indicated that the subject considered that the utterance was fluent, a score of 5 indicated detection of disfluency and intervening scores indicated uncertainty.

2.4. Results

In this analysis, only the 1-5 scores for the crucial point in the disfluent utterances (the first word of the restart) and the equivalent points in the control utterances are examined.

Subjects were able to give fluency judgements with considerable confidence. For disfluent utterances, they gave average scores of between 4 and 5 in the majority of cases (max = 50, min = 17, mean = 40.05); the controls received average scores of 1 or just over 1 (min = 10, max = 48, mean = 12.39, for all controls).

The differences between fluency judgements for critical points in disfluent utterances and the equivalent points in the controls were found to be significant (Friedman statistic by subjects = 38.2, df = 3, p < .001; by materials = 50.91, df = 3, p < .001).

There were 2 cases out of the total of 30 disfluencies where the total score for the disfluency judgement was lower than 30, indicating that on average subjects thought that the utterance might still be fluent. These scores were examined individually in Wilcoxon signed rank tests, comparing them with the scores for their fluent controls: there was still found to be a significant difference between the sets of scores, the scores for the disfluent items being higher than for their fluent controls (first case: n=6, W=0, p<.025; second case: n=7, W=0, p<.01).

2.5. Discussion

The subjects gave high scores of between 4 and 5 in the majority of cases where disfluency had occurred and low scores of between 1 and 2 where there was no disfluency, thus supporting the hypothesis that disfluency can be recognised by the offset of the first word after disfluent interruption.

3. EXPERIMENT TWO

3.1. Introduction

This experiment was designed to test the hypothesis that an editing signal at the interruption point prior to the continuation enables listeners to detect disfluency.

3.2. Materials

The materials used in this experiment were identical to those used in the first.

3.3. Subjects and Procedure

There were 20 subjects, as in the first experiment.

The procedure was the same as that in the first experiment except that the disfluency recognition task differed: subjects were asked to use the 1-5 scale to say whether they thought that, on the basis of what they had heard, the utterance would continue fluently or disfluently. Thorough explanations and practice sessions preceded the experiment.

3.4. Results

In this analysis, the critical point in the utterance is the word-gate prior to the restart.

Subjects showed less confidence in their fluency judgements than in the first experiment. They gave average scores of between 2 and 3 for the critical point in disfluent utterances (max = 3.7, min = 1.3, mean = 2.55); the average scores for the equivalent point in the controls were of 1 or just over 1 in most cases (min = 1.0, max = 3.7, mean = 1.9, for all controls).

The differences between fluency judgements for critical points in disfluent utterances and the equivalent points in the controls were found to be significant (Friedman statistic by subjects = 34.62, df = 3, p < .001; by materials = 21.77, df = 3, p < .001).

To examine the results for individual test items, Wilcoxon signed rank tests were performed, comparing scores for the spontaneous disfluent condition with those for the spontaneous fluent condition. The results of these tests show that the scores for the disfluent condition were significantly higher than those for the fluent condition in only 12 of the 30 cases (p<.05), the difference in scores was insignificant in 15 cases and the difference was significantly higher for the fluent condition in 3 cases.

3.5. Discussion

The results show that the hypothesis is only supported by a minority, 12, of the 30 test items. Of these 12, only 9 have average scores of 3 or over and the maximum is 3.7, which should indicate that subjects had a slight feeling that disfluency was about to occur.

A reexamination of the materials to search for any phonetic cues which may have caused higher scores reveals that the 12 test items for which the total scores were 30 or over fall into one of two main categories: words which are interrupted suddenly (incomplete words); words which are lengthened and/or followed by a pause and/or weakly offset or an inbreath. The majority of the other test items consist of complete words with no pause before the continuation.

The analyses suggest that listeners made use of cut-offs and hesitation phenomena, where they were present, in detecting oncoming repairs, but in the majority of cases, where such cues were not present, they were unable to detect imminent disfluency.

4. CONCLUSION

The experiments reported in this paper show that disfluency can usually be detected by the end of the first word following the interruption and do not support the hypothesis that listeners perceive and make use of a phonetically identifiable editing signal placed immediately prior to the onset of the continuation. Subjects only indicated that they detected oncoming repairs in a minority of cases. In the majority of cases, they appeared to make use of cues within the first word of the repair.

Further experiments are under way to determine more precisely where listeners can detect disfluency and to examine the contribution of prosodic cues to the perception of disfluency. It is suggested that rhythmic and intonational information plays a vital role in alerting listeners to the presence of disfluency, rather than a discrete phonetic editing signal.

7. REFERENCES


**ABSTRACT**

This study investigates whether individual variability in the categorisation of a voiced-voiceless speech contrast is related to the stimulus type used in the perceptual experiment. Continuous stimuli constructed using copy-syntheses, computer-edited natural tokens and stylised syntheses were used. Categorisation of reduced-cue continua was also examined for the copy-synthesised and natural-edited continua. Greater variability was generally found in the labelling of copy-synthesised continua.

1. **INTRODUCTION**

An initial study on the perception of initial stop place contrasts [1] has shown that individuals may vary greatly in the extent to which they are affected by the neutralisation of specific cues to these contrasts. Greater individual variability was found in the labelling of stops in an /æ/ environment than in an /e/ vowel environment and in the perception of complex syntheses, copied from a natural utterance, than that of more stylised syntheses.

The aim of the present study was to assess whether the variability was related to stimulus type, by controlling vowel environment. Stimulus types used included computer-edited natural speech, high-quality copy-synthesis based on the same natural tokens, and a highly stylised synthesised continuum created at the Haskins Laboratories [2].

2. **STIMULI**

The natural-edited and copy-synthesis versions were presented in three conditions:

a. **full-cue** (VOT and F1 cutback): change in VOT from -20 ms to +70 ms in nine steps and change in the first formant frequency.

b. **PaVOT:** same change in VOT with, at vowel onset, formant frequencies characteristic of [ba] (rising F1 onset throughout).

c. **PaVOT:** same change in VOT, with, at vowel onset, formant frequencies characteristic of [p] (flat F1 onset throughout).

2.1. **Natural edited stimulus continua**

Recording were made of tokens of /p/ and /ba/ produced by an adult English male speaker. Two tokens were chosen which were characterised by clear formant patterns and a regular fundamental frequency trace of around 100 Hz to facilitate editing in 10 ms steps. The creation of natural-edited stimulus continua was done on a mini-computer using a "cut and paste" technique. For the BaVOT continuum, the vowel portion from the [ba] token was appended to the burst and aspiration portion from the [p] token. For the creation of stimuli with VOT between 70 ms and 5 ms, the aspiration was progressively deleted from the vowel and in 10 ms steps. For stimuli with negative VOTs, the preceding portion was cut out from the [ba] stimulus and appended to the initial burst. The same process was carried out for the PaVOT continuum, except that, in this case, both the burst/ aspiration and vowel portions were taken from the [p] token. In the "full-cue" continuum, for steps with positive VOTs, initial cycles of the vowel portion were deleted as VOT increased to create formant cutbacks in voiceless tokens.

2.2. **Synthetic stimuli**

The natural tokens used as a base for the natural-edited continua were analysed using a ten pole closed-phase LPC analysis to derive the formant frequencies. Amplitude control parameters were obtained using an FFT analysis [3]. A first synthesis through a 4 kHz bandwidth, software parallel formant synthesizer was performed. Further modifications to the synthesis were then made on the basis of comparisons between the natural and synthetic spectra on a Kay digital spectrum graph until a close match was obtained. Analogous conditions to the ones created for the natural-edited speech were prepared. For more details on stimulus preparation, see [4].

The stylised synthetic Haskins continuum was presented in the full-cue condition only. The VOT range used was the same as above.

3. **SUBJECTS**

Subjects were 18 paid volunteers with normal hearing as defined by average thresholds of 10 dB HL or better, from .25 to 8 kHz. The listeners ranged in age from 18 to 29 years (mean: 20.7 years) and had no previous listening experience of synthetic speech.

4. **TEST PROCEDURE**

Stimuli were presented in the form of two-alternative forced-choice identification tests over four sessions. At each session, seven tests were presented. Each consisted of ten tokens repeated randomly eight times. Stimuli were presented at a comfortable listening level through headphones.

5. **RESULTS**

A statistical approach based on generalized linear models (GLMs) fit by maximum likelihood estimation was used to determine the extent to which performance varied across different test conditions. This technique, analogous to Analysis of Variance, was used as it is especially tailored to the analysis of multi-variate data involving binary responses (for a more detailed description, see [11]).

Using GLM, phoneme boundary and gradient measures were derived from the best fit cumulative normal to the four repetitions of each test condition for each of the 18 subjects (Fig. 1). A mean VOT phoneme boundary value was then derived for each of the three "full cue" conditions. For the mean boundary obtained were 13.5 ms (s.e. 5.1) for the natural edited condition, 13.3 ms (s.e. 6.2) for the stylised synthetic condition and 22.4 ms (s.e. 5.0) for the copy-synthesis condition. The mean gradient values obtained were -2.492 (s.e. 1.865) for the natural edited condition, 2.604 (s.e. 1.997) for the stylised synthesis condition, and -1.289 (s.e. 0.784) for the copy-synthesis condition. Highly similar phoneme boundary and gradient values were therefore obtained for the stylised syntheses and natural-edited stimuli. The copy-synthesis condition was less sharply labelled and showed a shift in boundary.

The next step of the analysis was to investigate difference in labelling between conditions for individual subjects. For each subject, the condition deviances, which are quantitative, statistically interpretable, measures of the extent to which subjects change their labelling behaviour across conditions (see [1]) were calculated. Labelling of the stylised synthetic condition was...
compared with labelling of the other full-cue conditions. 83% of subjects showed significant deviations at the 0.001 level (deviances greater than 26.1) between the copy-synthesised and stylised synthesis continua. Significant deviations were only found for 44% of subjects when the natural edited and stylised synthesis stimuli were compared and the range of deviations obtained (8.9 to 61.9) was generally smaller than in the first comparison (22.6 to 174.6).

Next, the effects of cue reduction on phoneme boundary and gradient for copy-synthesised and natural-edited continua were examined. For the natural edited stimuli, the mean phoneme boundary increased from a value of 13.52 ms for the full-cue condition, to 16.89 ms (s.e. 6.14) for the Ba/VOT condition and decreased to 0.47 ms (s.e. 11.73) for the Pa/VOT condition. For the copy-synthesised stimuli, the shift was from 22.41 ms for the full-cue to 25.04 ms (s.e. 5.53) for the Ba/VOT and 10.1 ms (s.e. 15.4) for the Pa/VOT condition.

Condition deviations were again calculated to compare labelling for the full-cue condition and the two reduced-cue conditions for individual listeners. For the natural edited range, very few listeners (11%) showed a significant deviation (p<0.001) between the full-cue and Ba/VOT condition. For the copy-synthesised stimuli, a greater number of listeners (33%) showed such an effect. Greater differences in labelling were found between the full-cue and Pa/VOT conditions. Generally greater individual variability in the labelling of this reduced cue condition was obtained, showing that some listeners were more greatly affected by changes in the spectral characteristics than others (Fig. 2). With the natural edited stimuli, all listeners showed a significant deviation between the two conditions with condition deviations ranging from 58.7 to 201.4, while, with the copy-synthesised stimuli, only 72% showed such an effect (deviances ranging from 9.4 to 240.2).

6. DISCUSSION
When full-cue ranges were presented, more similar results were obtained for natural-edited and highly stylised Haskins synthetic continua than for a copy-synthesised continuum based on parameters measured from the same natural tokens. One explanation might be that, in the Haskins continuum, the unnaturalness of the highly stylised stimuli is compensated by the clear enhancement of the cues which are present. With the copy-synthesised stimuli, listeners are having to deal with a complex set of patterns which may also contain slight inaccuracies in terms of formant bandwidth values and intensity relations for example. Certain listeners, especially in reduced-cue conditions, may be more sensitive to these inaccuracies and as a result, show greater variability in categorisation.

When looking at the effect of cue reduction, it was found that the lack of an appropriate F1 onset with short VOT (Pa/VOT condition), generally led to a smaller number of “voiced” responses, showing the importance of spectral cues to the voicing contrast. For both stimulus types, individual listeners varied in the extent to which they were affected by the spectral cue to the voicing contrast as shown by large differences in condition deviation measures obtained. However, more homogeneous results were obtained with natural-edited stimuli than with copy-synthesised stimuli.

7. REFERENCES
PERCEPTUAL SPACES OF THE RUSSIAN VOWELS

G.N. Lebedeva

Institute of Chemistry and Technology, Ivanovo, USSR

ABSTRACT

The aim of this work is to analyse the perception of non-native vowels by the native speakers of Russian. The main tasks are: 1) the establishment of perceptual spaces of the Russian vowels under different conditions; when identifying non-native vowels a) isolated from the phonetic context; b) in CV and VC syllables; 2) the definition of the number of the distinguished vowels. The results allow us to maintain that untrained speakers of Russian are able to identify reliably 8 non-native vowels and to distinguish 18 vowels.

1. INTRODUCTION

It's well-known that the number of Russian vowel allophones which the native speakers of Russian are able to distinguish is much greater (n=10) than the number of the Russian vowel phonemes is (n=6)[1]. Numerous experimental studies of late assure us of the fact that Russian listeners possess a highly developed system of perception of phonetic features of vowels. One of the latest works in this field is that fulfilled by Tchernova and colleagues[2]. The authors investigated the perception of 20 cardinal vowels by the untrained speakers of Russian. In the first experiment the listeners were asked to identify all the cardinal vowels using only 10 symbols (the letters of the Russian alphabet) as possible answers. It was revealed that listeners were able to distinguish about 17 vowels among the 20. In the second experiment the listeners were preliminarily taught to transcription, then they listened to a vowel "sample" marked by a certain transcription sign. The listeners were able to discriminate all the number of identified vowels however increased but little (n=9-10). The problems raised in such works seem to be very actual both from the viewpoint of establishing the correlation between the perceptual and the phonological units, and from the viewpoint of elaboration of the strategy of foreign language teaching.

2. PROCEDURE

At different periods of time three groups of untrained speakers of Russian were asked to identify the vowels of English, Spanish and German. English and Spanish vowels were isolated from the words within which they were pronounced, the German vowels were presented for identification within CV and VC syllables. The number of answers given by the Russian speakers, though the total number being 36. As it's seen from the Table, the listeners use the Russian letter (3) more frequently than other letter-symbols to mark the vowels they were presented to. The space includes 6 vowels: /æ/ , /a/ , /ɔ/ , /e/ , /i/ , /u/. The space of the Russian (4) and (5) include 5 vowels: /a/ , /ø/ , /ɛ/ , /ɔ/ , /u/ , /æ/ . The most narrow are the spaces of (6) and (7) including only one vowel each: /æ/ , /a/ , /i/ , /u/ . The vowel system of Spanish very much resembles that of Russian (as far as the number of vowel phonemes is concerned). All these facts are of great interest from the viewpoint of the study of the mechanisms of phonological hearing.

3. THE CHOICE OF THE LANGUAGES

The choice of the languages under study was not accidental. It was conditioned by the facts that, on the one hand, the vowel systems of English and German contain practically all the possible types of vowels, on the other hand, the vowel system of Spanish very much resembles that of Russian (as far as the number of vowel phonemes is concerned). All these facts are of great interest from the viewpoint of the study of the mechanisms of phonological hearing.

4. RESULTS

The results of the identification test are presented in Table 1. In its vertical column the table contains only the number of answers given by the Russian speakers, though the total number being 36. As it's seen from the Table, the listeners use the Russian letter (3) more frequently than other letter-symbols to mark the vowels they were presented to. The space includes 6 vowels: /æ/ , /a/ , /ɔ/ , /e/ , /i/ , /u/. The space of the Russian (4) and (5) include 5 vowels: /a/ , /ø/ , /ɛ/ , /ɔ/ , /u/ , /æ/ . The most narrow are the spaces of (6) and (7) including only one vowel each: /æ/ , /a/ , /i/ , /u/ . The vowel system of Spanish very much resembles that of Russian (as far as the number of vowel phonemes is concerned). All these facts are of great interest from the viewpoint of the study of the mechanisms of phonological hearing.

Fig. 1. The distribution of the listeners' answers (by the y-axis) on the vowels which give the most reliable identification. within CV and VC syllables. The number of answers given by the Russian speakers, though the total number being 36. As it's seen from the Table, the listeners use the Russian letter (3) more frequently than other letter-symbols to mark the vowels they were presented to. The space includes 6 vowels: /æ/ , /a/ , /ɔ/ , /e/ , /i/ , /u/. The space of the Russian (4) and (5) include 5 vowels: /a/ , /ø/ , /ɛ/ , /ɔ/ , /u/ , /æ/ . The most narrow are the spaces of (6) and (7) including only one vowel each: /æ/ , /a/ , /i/ , /u/ . The vowel system of Spanish very much resembles that of Russian (as far as the number of vowel phonemes is concerned). All these facts are of great interest from the viewpoint of the study of the mechanisms of phonological hearing.

Table 1. The distribution of the listeners' answers (by the y-axis) on the vowels which give the most reliable identification within CV and VC syllables. The number of answers given by the Russian speakers, though the total number being 36. As it's seen from the Table, the listeners use the Russian letter (3) more frequently than other letter-symbols to mark the vowels they were presented to. The space includes 6 vowels: /æ/ , /a/ , /ɔ/ , /e/ , /i/ , /u/. The space of the Russian (4) and (5) include 5 vowels: /a/ , /ø/ , /ɛ/ , /ɔ/ , /u/ , /æ/ . The most narrow are the spaces of (6) and (7) including only one vowel each: /æ/ , /a/ , /i/ , /u/ . The vowel system of Spanish very much resembles that of Russian (as far as the number of vowel phonemes is concerned). All these facts are of great interest from the viewpoint of the study of the mechanisms of phonological hearing.

Fig. 1. The distribution of the listeners' answers (by the y-axis) on the vowels which give the most reliable identification. within CV and VC syllables. The number of answers given by the Russian speakers, though the total number being 36. As it's seen from the Table, the listeners use the Russian letter (3) more frequently than other letter-symbols to mark the vowels they were presented to. The space includes 6 vowels: /æ/ , /a/ , /ɔ/ , /e/ , /i/ , /u/. The space of the Russian (4) and (5) include 5 vowels: /a/ , /ø/ , /ɛ/ , /ɔ/ , /u/ , /æ/ . The most narrow are the spaces of (6) and (7) including only one vowel each: /æ/ , /a/ , /i/ , /u/ . The vowel system of Spanish very much resembles that of Russian (as far as the number of vowel phonemes is concerned). All these facts are of great interest from the viewpoint of the study of the mechanisms of phonological hearing.

Table 1. The distribution of the listeners' answers (by the y-axis) on the vowels which give the most reliable identification within CV and VC syllables. The number of answers given by the Russian speakers, though the total number being 36. As it's seen from the Table, the listeners use the Russian letter (3) more frequently than other letter-symbols to mark the vowels they were presented to. The space includes 6 vowels: /æ/ , /a/ , /ɔ/ , /e/ , /i/ , /u/. The space of the Russian (4) and (5) include 5 vowels: /a/ , /ø/ , /ɛ/ , /ɔ/ , /u/ , /æ/ . The most narrow are the spaces of (6) and (7) including only one vowel each: /æ/ , /a/ , /i/ , /u/ . The vowel system of Spanish very much resembles that of Russian (as far as the number of vowel phonemes is concerned). All these facts are of great interest from the viewpoint of the study of the mechanisms of phonological hearing.
estimated by means of $X^2$ criteria. The comparison of the distribution shows that the listeners are able to distinguish not 8 but at least 14 vowels. As it is seen, 4 vowels are placed higher than the critical meaning of the $X^2$ criteria i.e. Thus, these vowels are very well distinguished by the listeners too. The vowels placed below the critical meaning of the criteria on one vertical line with the vowels marked (·) (see Fig.1), to all appearance, seem to be identical for the Russian listeners as far as their phonetic features are concerned. Fig.1 gives the opportunity to represent, firstly, the width of the perceptual boundaries of the Russian vowels, and secondly, the remoteness of non-native vowels from the centre formed by the native vowel in the name's consciousness of the Russian speakers. Let's analyse another group. While identifying these vowels the listeners do not take unanimous decisions. The vowels /ɛ:/, /æ:/, /ɛn/ (see Table 1). The task which the listeners had to fulfill was undoubtedly very difficult: to place the vowel they heard into a certain sphere of a perceptual space formed in their memory by the native sounds and to correlate the articulation of the unknown vowel stimuli. Let's consider the vowels which differ only in one step of openness articulations of vowels to each other. Thus, mistakes to within one step of openness are considered to be possible. Then the identification of some of the above mentioned vowels improves. For example, the English /æ/ and /ɛ:/ are identified mostly as /ɛ:/ (70% and 60% of all the answers respectively). Comparison of the answers' distributions shows that these two vowels form the perceptual sphere for the speakers of Russian and they may be placed in the space of /æ/ in Fig.1. German /i:/ is identified in the 50% of all the answers as /ɪ:/ and in the 50% as /ɛ:/ Thus, it may be placed in the space of /ɛ:/ on Fig.1. English /i:/ and Spanish /o:/ are identified as the Russian /o:/ (approximately 50% of all the answers corresponding). The analysis of the distributions shows that these vowels stay close to each other as far as their phonetic properties are concerned and they form a common area which can be placed in the space of /o:/ on Fig.1. While identifying the German /y/, /ʊ:/, /pː/; the Russian listeners use from 6 to 14 symbols and combinations of symbols. These are mostly the combinations of a front close vowel with a back rounded vowel. This fact testifies to the phonological characteristics of the operation: the mechanism of the front rounded vowel identification resembles that used by the native speakers of Russian when identifying the Russian vowel allomorphs in the position between or after palatalised consonants [2]. The analysis of the distribution of inadmissible answers (Fig.1) gives us: 1) the vowels which are placed in a perceptual space of a definite Russian vowel. Their number is 10 and they form vertically clearly distinguishable rows; 2) the vowels which are placed in a perceptual space, formed in a linguistic consciousness of the Russian speakers by several symbols: /uː/; /æː/; /ɛː/; /oː/; (see Table 1) these are not placed in a perceptual space ("alien to it"). These are the German front rounded vowels.

5. DISCUSSION

The results of the investigation allow us to maintain that in the case of a non-native vowel identification the Russian listeners are able to identify reliably 8 vowels. The number of distinguished vowels is equal to 14 (Fig.1) /æː/, /ɛː/, /æ/, /yː/, /ɪ:/, /ʊ:/, /pː/. All the vowels can be divided into 3 groups as far as their perceptual estimation by the native speakers of Russian is concerned; 1) the vowels which are placed reliably in a perceptual space of a definite Russian vowel. Their number is 10 and they form vertically clearly distinguishable rows; 2) the vowels which are placed in a perceptual space, formed in a linguistic consciousness of the Russian speakers by several symbols: /uː/; /æː/; /ɛː/; /oː/; (see Table 1) these are not placed in a perceptual space ("alien to it"). These are the German front rounded vowels.

6. REFERENCES

A CROSS-LINGUISTIC EXPERIMENTAL INVESTIGATION OF SYLLABLE STRUCTURE: SOME PRELIMINARY RESULTS

Bruce L. Derwing, Sook Whan Cho and H. Samuel Wang
U of Alberta, Canada Sogang U, Korea National Tsing Hua U, Taiwan

ABSTRACT
Prior research has shown that there is more to English syllables than a mere linear sequence of phonemic segments. The present research attempts to extend the use of techniques developed in the English investigations to the study of comparable phenomena in other languages of diverse types. A preliminary report is given on the status of sub-syllabic units in Taiwanese and Korean, together with some new findings on Korean syllable boundaries.

1. BACKGROUND
The experimental investigation of syllable structure began with the work of Treiman [1,2, etc.], who used a variety of string manipulation tasks (notably word-bending) to determine whether such hypothesized units as the onset, rime or codas were viable for English. Dow [3,4, etc.] continued this work, using primarily a unit-substitution task. Taken together, this research lends support to the idea that English syllables have an onset-rime or right-branching structure. Treiman & Danis [6] have recently extended this investigation to the question of syllable boundaries in English, putting such notions as the Maximal Onset Principle to experimental test. A chief purpose of the present study was to extend or adapt the methodologies developed in these English language investigations to other languages of diverse types, in order to explore the question of the generality of the findings.

2. SUB-SYLLABIC UNITS IN TAIWANESE AND KOREAN
2.1 Taiwanese
Since the initial attempts to apply Dow's unit-substitution task to Arabic, Blackfoot and Taiwanese proved impractical, it was decided to try a forced-choice version of Treiman's word-bending task that could be group administered. Since the main question of interest related to the direction of the primary boundary between the vowel and adjacent consonants, subjects were given two alternative 'blendings' of a pair of Taiwanese words, one which combined the onset of one with the rime of the other and a second which combined the head of one with the tail of the other, as illustrated by the following example: SAN1 + CIM1 \rightarrow (a) SIM1 (b) CIN1. (The numbers following each Taiwanese word indicate tone.) Also included on the test were several word pairs like the following, where both choices were of a single type: TA5 + PI5 \rightarrow (a) TIS5 (b) FAS5. By comparing the results on these items with the first group, we could assess whether there was a distinct preference for one type of blend over the other.

The forced-choice word-bending task was conducted in Taiwan in November 1990 and in January 1991, yielding 95 subjects in all. The results, however, revealed no distinct preference in favour of either onset-rime or head-coda blends, as responses to the 'choice' and 'non-choice' items were indistinguishable: in both cases responses were essentially random, except for a slight overall bias in favour of choosing the first response, regardless of type. This presumably means one of two things: (1) perhaps our subjects did not understand the nature of the task, or else were simply not able to perform it reliably under the conditions it was presented; (2) alternatively, perhaps the simple monosyllables of this language, involving no consonant clusters and very severe internal collocational constraints, are not readily analyzable by speakers into smaller units. This second interpretation is consistent with the results of Read et al. [7] from a related dialect, in which ordinary subjects (i.e., subjects not familiar with the pinyin alphabetic transcription scheme) proved unable to perform the simple task of replacing the initial consonant (onset) of a Mandarin word by another consonant; instead, their performance was highly parallel to that found by Morais et al. [8] in a similar task with literate Portuguese speakers. (See [9,10] for further discussion of problems with the notion of the phone as a universal unit of speech segmentation.)

2.2 Korean
The Korean language is of much interest to this investigation, as there are reasons to believe that syllables in this language reflect a head + coda structure rather than an onset + rime organization of English (i.e., unlike English, vowel nuclei in Korean seem to adhere more closely to preceding consonants than to following ones). Native speakers report this to be the case on the basis of their own intuitions, and even the standard orthography reflects a judgment of this kind. The syllable SAN (meaning 'mountain'), for example, is represented at two vertical levels, with Korean letters for SA placed on top and the letter for N placed below it, thus weighing a combination like 'SAN' rather than 'SAN'. In addition, Youn has recently conducted an informal word-bending production task, whose results did date support this analysis (see [11]). A Korean version of the forced-choice word-bending task is now under way to firm up these preliminary findings, but the results of this study are not yet available.

3. SYLLABLE BOUNDARIES IN ENGLISH AND KOREAN
3.1 English
Initial attempts to apply the Treiman & Danis (T&D) syllable-inversion task to other languages were generally unfruitful: less than 10% of our Arabic subjects, for example, were able to perform any inversion at all. When a problem of this nature emerged in the early stages of the Blackfoot investigation, it became clear that a new, simpler technique was going to have to be developed, one that would not require literacy skills to perform. (This was especially critical for Blackfoot, as few speakers know the orthographic system that has been developed only recently by linguists for that language.)

A new technique that worked involves what we call the 'pause-break' task. In this task subjects are asked to choose which of two or three alternative 'breakings' of a word sounded the 'most natural.' To illustrate for the English word MELON, for example, the following three alternatives were offered (where \(\textit{e}\) ... indicates the location of the pause):
(a) /\textit{m} \textit{el} \textit{on}/ (where \(\textit{e}\) is treated as the onset of the second syllable),
(b) /\textit{m} \textit{elin}/ (where \(\textit{e}\) is the coda of the first syllable), or
(c) /\textit{m} \textit{el} \textit{in}/ (where \(\textit{e}\) is an ambisyllabic). This was especially critical in the English students, all native speakers with little or no prior exposure to linguistics or phonetics. The main purpose of this pilot study was to evaluate whether the earlier T&D results, using more difficult tasks, could be replicated, and, as indicated in [5], the answer was in the affirmative. This new task has thus been adopted for testing in most of the languages in the project, but only the Korean data are available at this time.

3.2 Korean
In the Korean writing system (called hangul), letters are used for individual segments and written from left to right, much as in English, but, by utilizing the vertical dimension as already noted above, these letters are also grouped into syllable-sized "bundles." The hangul spelling of each Korean word thus makes a commitment as to the location of the syllable boundary which even literate speaker presumably knows. The purpose of the present investigation, therefore, was to establish whether any general preference could be found that was inde-
rudent of the orthographic norms.

In principle, we saw two possible ways to investigate this. One possible course of action, obviously, would be to carry out the study among illiterate speakers, who would not know the orthographic norms. The second approach, which could be more readily implemented, was to focus the investigation on homophones having a variable placement of the orthographic syllable boundary, as exemplified by the morphological structure of the words involved. The phonemic string MILI in standard Korean, for example, is ambiguously syllabified in the orthography as MIL/I (when it means 'in advance') or as MIL/I (when it means 'wheat + noon'), where a slash is used to show the location of the break between the syllabified 'hangul' packages. For subjects who were given the meanings of the Korean words in the oral presentation used in our study, we expected a close conformity to the orthographic norms. For the other group, however, who were not given the meanings, we saw a possibility for some general phonological preferences to emerge.

The first round of Korean data was collected in October 1990 in Seoul, when two groups totaling 117 subjects were presented with six items similar to the one above, as well as a number of supplementary items selected to test cases mostly involving intervocalic tense consonants or consonant clusters. All subjects were undergraduate students in the Department of English at Sogang University, the majority of whom grew up in the general Seoul area. The results were as follows:

1. The clearest cases involved single consonants that are restricted phonetically to syllable-initial position, such as /S/ (as in SONG/SO [99]) or to syllable-final position, such as /G/ (as in FANG-GI [1.00]).
2. The results were also very clear for consonant clusters, with the preferred break occurring between them. This result was virtually unanimous if this break corresponded with the spelling (as in CHENG/ISO [99] and KUK/KU [98]), but remained the majority choice even when the orthography put the break after the second consonant (e.g., AN/GIA [74] and KAP-SD [66]).
3. For tense consonants (written as geminates) the results were also fairly clear, with the preferred break once again after the vowel in spelling-supported cases (e.g., A-PPA [99] and KA-CCA [79]), but with a major shift to the spelling break if it occurred after the consonant (e.g., MU/KK/E [45] and KA-SS/E [32]).
4. In the crucial orthographically ambiguous strings, which mostly involved single intervocalic consonants, the preferred break position was immediately after the vowel; however, as shown in the summary of these results below, the size of the plurality varied considerably as a function of consonant-type. *Note that two figures are given for these words: the first shows the proportion of subjects who broke the words at the hyphen under the 'no meaning' or 'ambiguous string condition, while the second shows the result when the meanings were supplied.*

**NOTES**

1. The research reported here was supported by a research grant from the Social Sciences and Humanities Research Council of Canada (No. 410-88-0266), awarded to the first author. The authors also wish to express their deep thanks to Y.B. Youn (Sogang University), whose aid was indispensable to this project, and to T.M. Nearey for his technical assistance.
2. More recent work has suggested an alternative interpretation that is less hard and fast (see [5], in this volume).
3. In all of these examples, a hyphen is used to show the judged syllable break and a slash (\/) to show where the break occurs in the spelling; if both breaks coincide, the composite symbol '/' is used. The numbers indicate the proportion of subjects who chose to break the words at the place marked by the hyphen.
4. Note that the suggested hierarchy is much the same as that found for English (see [5], in this volume), except that the linkages in Korean, as expected, are the following vowel, rather than to the preceding one.

**REFERENCES**

LA PHONÉTISATION DU CASTILLAN
CABRERA C., CONTINI M. et BOÉ L.-I.
Institut de la Communication Parlée, URA CNRS n° 368
Grenoble, France

ABSTRACT
Our project was the establishment of a grammar for the automatic phonétisation of Spanish. By examining a lexicon of 63,000 words and systematically examining their transcriptions, we formulated a large body of rules. In a next step, we will use this knowledge for a text-to-speech synthesis application. We constituted a data base of 2500 words. The resulting system gives a correct phonétisation of 98% of the original lexicon. We here present the analysis method used on this large lexicon, as well as a selection of the rules derived.

1. INTRODUCTION
La phonétisation automatique consiste à passer d'une chaîne orthographique quelconque à une chaîne phonétique. Cette transcription en A.P.I. ou dans un autre code relève du domaine de la description linguistique et peut être utilisée dans une application telle que la synthèse de la parole. Ses intérêts sont multiples et apparaissent de plus en plus comme l'étape nécessaire pour l'établissement du dialogue homme-machine. Sur le plan hispanique, ce champ d'étude a déjà fait l'objet de récents travaux [1].

Résultat d'une collaboration étroite entre linguistes et informaticiens, l'outil de phonétisation multilingue qu'est TOPH [2], défini dans le cadre de la synthèse à partir du texte, présente l'avantage pour le linguiste de formaliser facilement sa connaissance. Conçue comme un module adaptable à chaque langue orthographique visée en l'occurrence le français, l'allemand et l'italien, cet outil a donné lieu au développement de grammaires de transcription pour chacune de ces langues. Cette étude se veut une description linguistique des phénomènes de phonétisation mais une étape ultérieure consistera à l'intégration de ces connaissances dans un système de synthèse (SYNTALIT).

Notre contribution consiste en l'établissement d'une grammaire de règles de transcription orthographique-phonétique utilisant le formalisme TOPH pour le castillan normatif (prononciation de l'espagnol madrilène cultivé).

2. PRÉSENTATION DE TOPH
Formalisation de grammaires de transcription, TOPH a été réalisé afin de proposer une description concise des phénomènes de phonétisation. Le logiciel élaboré s'articule autour des éléments synoptiques suivants :
- L'unité linguistique sélectionnée est la chaîne graphémique
- Déclaration d'ensembles de natures différentes à savoir les ensembles linguistiques et les lexiques d'exceptions.
- Le linguiste formalise son raisonnement sous la forme d'une grammaire déterministe (à une quelconque sous-chaîne d'un mot correspond une seule transcription) de règles de réécritures contextuelles.
- Ordonnées du particulier au général, les règles sont regroupées par classes, avec un ordre local pour chaque règle, défini par son ordre d'écriture.

Possibilité d'inserer de commentaires dans la grammaire bornés par ! L'intérêt de TOPH réside dans l'accès à des tracés de réalisation des règles sollicitées de même qu'à des résultats statistiques sur ces dernières.

3. GRAMMAIRE DU CASTILLAN
La grammaire a été élaborée sur la base de 65000 entrées lexicales issues du dictionnaire SGEL [3] dont la particularité, ouvre les transcriptions attachées à chaque enreg, réside dans l'introduction de nombreux emprunts (anglicismes en majorité) plus ou moins assimilés au phonétisme du castillan. A l'aide d'un ensemble de règles la correspondance phonétique de chaque graphème est définie en tenant compte de toutes ses distributions possibles. L'apport constant de termes nouveaux auxquels une langue naturelle est soumise nécessitera une mise à jour régulière de notre grammaire. Ceci pose évidemment le problème de la pertinence des lexiques liée à leur actualisation.

Pour la prononciation standard du castillan nous nous référions à des ouvrages spécialisés [4], [5], [6]. Nous nous appuyons en outre sur le dictionnaire SGEL (mentionné précédemment) à partir duquel nous avons dressé des listes d'exceptions pour chacune des 29 lettres de la langue. Ces listes contiennent toutes les réalisations phonétiques des lettres supplémentaires ainsi que par rapport aux règles mentionnées dans les travaux déjà signalés dans le but de répertorier toutes les occurrences allophoniques pour une chaîne graphémique donnée afin de construire une grammaire de phonétisation la plus complète possible. Après ce premier travail d'identification et de synthèse, nous nous sommes attachés à l'édition et à la codification de la grammaire proprement dite pour laquelle nous avons déclaré les éléments décrits ci-après :

- 12 ensembles répartis en ensembles linguistiques par exemple :
  a) "semi-consonnes" = (x, w)
  b) séparateur de mots
  "w" = (s, t, z, ı, ı, ı)
  c) "except" : l" = (articulad, angular, unvove, axic, auricular, atomic, ocnicita, odegradable, odegradation, odisamica, ofisica, ograf, ográfico, ógrafa, ologia, ológico, ologo, oluminiscencia, omaia, omeánica, ometria, opsia, oquimico, osfera, osftisites, oterapia, ótico, otila, oirospism, óxido, al, ofita, oszoos, alin, aino, ant, oisible, oide, able, abild, enio, enal, edro, ásico, ar, ángulo)

Cet ensemble d'exceptions nous permet d'écrire la règle :
"(w)+b,br,hr,rv]+w ("exception : l") = [1]
sachant que la règle générale (majoritaire)
4. RÉSULTATS
A la lumière des résultats, plusieurs remarques s'imposent. Il apparaît que si l'on ne considère que les règles de prononciation circonscrites aux phénomènes réguliers, il est d'autant plus facile de tenir compte des exceptions ou des emprunts, la phonétisation du castillan se résume à une soixantaine de règles élémentaires. A titre illustratif, nous nous limitons au cas du graphème “g”. Alors que ce graphème est communément défini comme se réalisant selon 3 allophones, il s'enrichit de nombreuses réalisations lorsque nous étendons la grammaire à l'étude des emprunts et autres exceptions (entrecroisement).
Ainsi si l'on considère le trait régulier, un graphème comme “g” sera traité par 3 règles:

\[ +g + (c,i) = [x] \]
\[ ("#", n) +g+ = [g] \]
\[ +g+ = [y] \]

En revanche, il en fallait 19 si l'on tient compte, par ailleurs, des emprunts:

\[ ("#"-gon,buld) +g+ ("#") = [g] \]
\[ ("#"-ziga,icehr,bas) +g+ ("#") = [x] \]
\[ ("#"-ban,campia,smokin,bumeran,boom, ran,rin,puddin,pin,pon,parkin,marketin, gon,dumpin,dopin) +g+ ("#") = [] \]
\[ ("#"-tun) +g+ (steno) = [] \]
\[ ("#"-témín) +g+ (ton) = [] \]
\[ ("#"-gan) +g+ (stes-simo, m) = [] \]
\[ ("#"-copyst,br) +g+ (tn) = [] \]
\[ ("#"-neglig) +g+ (e) = [] \]
\[ ("#"-sutra) +g+ (is++m,t) = [y] \]
\[ ("#"-be) +g+ (elis+nismo) = [y] \]
\[ ("#"-per) +g+ (ola) = [g] \]

Les règles ont été testées sur une base de données conséquente et notamment un dictionnaire de 2500 entités, implanté sur HYPERCARD (Macintosh) contenant des formes orthographiques et phonétiques de référence. Actuellement 484 règles et 3 ensembles d'exceptions permettent de phonétiser automatiquement ce corpus. Nous obtenons un taux de succès de 93%.

5. CONCLUSION
Dépourvu d'homographes hétérophones, le castillan s'avère être une langue relativement régulière quant à un processus de phonétisation. Néanmoins si l'on considère la manière dont elle intègre les emprunts, nous conçois que ces apports lexicaux désorganisent quelque peu le phonétisme de cette dernière. De plus, si le castillan ne suit pas les règles de prononciation standard, elle conserve deux phonétismes possibles pour une même unité lexicales, une se fondant sur le phonétisme castillan et l'autre conservant les traits de la langue d'origine (bridge, chauvinismo).

RÉFÉRENCES
LANGUAGE SPECIFIC PATTERNS OF PROSODIC AND SEGMENTAL STRUCTURES IN SWEDISH, FRENCH AND ENGLISH.

Gunnar Fant, Anita Kruckenberg and Lennart Nord

Department of Speech Communication and Music Acoustics, KTH, Box 700 14, S-100 44 STOCKHOLM, SWEDEN.

Phone 46 8 790 7872, Fax 46 8 790 7834

ABSTRACT

This is a study of temporal patterns of stress in Swedish, English and French, focusing on durations of syllables and phonemes in stressed and unstressed positions. In French we note a finite amount of stress induced segmental lengthening at phrase initial locations which is less prominent than phrase final prepause lengthening and also smaller than in Swedish and English. If compared on the basis of the same number of phonemes per syllable the stress induced lengthening is less in French than in the two other languages. These results are interpreted within the concept of "stress timing" versus "syllable timing".

1. INTRODUCTION

The main purpose of our presentation is to report on some experiments on the realization of stress patterns. We have recently extended our studies of Swedish prose reading [4] to a pilot study of French and English [5]. A primary object has been durational structures. How does stress influence the duration of syllables and individual speech sounds? To what extent will language specific differences in syllable complexity influence overall durations of stressed and unstressed syllables?

Our results contribute somewhat to the perspective of "stress timing" versus "syllable timing". We have results from a small pilot study of a Swedish text translated into French and English. A few remarks about terminology may be needed. In French phonetics [7] the term "stress" is often avoided and is replaced by the partial term "accent", e.g. in connection with so called "accent d'insistance", indicating a marked accent usually falling in a syllable preceding the one that would otherwise have been expected to receive some degree of prominence. In French, the phrase and sentence groups, outlined by the intonational pattern and further marked by group final lengthening, is considered primary. In addition, however, there exists - just as in English and Swedish but less apparent - a subdivision of a phrase into smaller units around content words that are mainly marked by local F0 contours. This is what Delattre refers to as "mote continuations" [3]. One outcome of our study is to verify the existence of these prosodic word accents, and to quantify their small but usually finite durational correlates. We have found it profitable to make a general distinction between these minor accents and those which are followed by a pause. Their durational patterns are systematically different.

2. RESULTS

Our studies confirm this general view. In all three languages, stressed or accented syllables display a prolonged duration. In French, the stressed induced syllable lengthening is not limited to phrase final prepause lengthening. The phrase internal minor accentuations are associated with an increase of the order of 50 ms, compared to 100-150 ms for English and Swedish. In French the durational component is often negligible, whilst a typical slow rise of F0 followed by a faster resetting constitutes the remaining cue. Prepause lengthening was found to be greater in both Swedish and English compared to French.

A closer view of stress induced lengthening within a stressed syllable reveals characteristic differences. In all three languages, prepause lengthening affects phoneme durations in essentially inverse proportion to their distance to the boundary. In French, this pattern contrasts drastically to that of the internal minor accents, where consonants after the stressed vowel do not appear to receive any stressed induced lengthening. As shown in Fig. 1, the lengthening profiles within stressed syllables in nonterminal position are different for French, English and Swedish, with an overweight on consonants following the vowel in Swedish and consonants preceding the vowel in French, whereas in English the profile is more symmetrical.

We shall now look more closely into average stressed and unstressed syllable durations in the three languages. Following traditional definitions of syllables and excluding prepause stresses, we found rather similar values for unstressed syllables, 125 ms for Swedish, 140 ms for English and 130 ms for French. The corresponding values for stressed syllables were 290 ms for Swedish, 300 ms for English and 220 ms only for French.

However, we may argue to what extent these differences depend on syllable complexity. For unstressed syllables we find 2.1 phonemes per syllable for French and 2.3 for French and English. In French the particular distribution is very much dominated by two-phoneme and three-phoneme CV type. An apparent difference exists with respect to stressed syllables. In our text we found an average of 3.0 phonemes per syllable.

for Swedish, 3.1 for English and 2.3 for French.

Do these differences fully explain the durational data? The answer is no. Our procedure for the test is more fully described in [5]. It accounts for plotting syllable durations against stressed syllables. For Swedish unstressed syllables we find

\[ d = 10 + 50m \]

where \( d \) is the syllable duration and \( m \) the average number of phonemes. For English and French we found somewhat larger values for \( m \) greater than 2. For Swedish stressed syllables we obtained \( D = 57 + 77m \).

\[ d = 81 + 32m \]

\[ D = 81 + 32m \]

(2) The result was similar for English, whilst for French we noted a best fit in terms of

\[ D = 81 + 32m \]

(3) Now, if we compare the Swedish and the French data with respect to the same number of phonemes, e.g. \( m = 3 \), we find \( D = 290 \) ms for Swedish and 235 ms for French. This analysis reflects a true stress induced difference.

We shall now take a more detailed view of the differences between stressed and unstressed syllables in the three languages. Fig. 2 shows successive syllables within a long sentence in English, French and Swedish. Here the ordinate is the difference in duration between a syllable and an unstressed reference, determined as the sum of average un-
3. STRESS TIMING VERSUS SYLLABLE TIMING

"Stress timing" versus "syllable timing" are concepts frequently used in language descriptions. The stringency and relevance of these terms, originally coined by Pike [6] and promoted by Abercrombie [1], have often been questioned. What evidence do we have for referring to Swedish and English as "stress timed" and French as "syllable timed"? The initial postulate concerning stress timing was a constancy of interstress intervals irrespective of the number of syllables contained. Since long, this extreme postulate has been refuted [2]. Here follows a condensed summary of our earlier discussion on this issue [5]:

1. Even though weak isochrony tendencies are found in English and Swedish, they do not seem to be of a sufficient perceptual salience to serve as a basis for a theory of stress timing versus syllable timing.

2. Most content words receive some degree of accentuation also in French, which potentially could constitute a basis for stress timing just as in English or Swedish. However, in French the phrase internal stresses are less apparent, whilst the regularity of the succession of syllables becomes dominant.

To sum up, we have found that the smaller contrast between stressed and unstressed syllable durations in French compared to English and Swedish is both a matter of a smaller contrast in syllable complexity and a lower degree of stress induced lengthening. In addition, the relative precision and low degree of vowel reduction in unstressed syllables in French reduces the stressed/unstressed contrast. Another argument in the same direction is the relatively moderate F0 span of nonterminal stresses compared to the more dominant prepause contours. In our view the main arguments for referring to French as "syllable timed" and Swedish and English as "stress timed" are the following:

The stress timing is not a matter of physical isochrony of interstress intervals, but a perceptual dominance of heavy syllables, the succession of which is sensed as quasi-periodical. A language is sensed as syllable timed, when these stress cues, including contrasts in syllable complexity and precision are reduced.

ACKNOWLEDGEMENTS

These studies have been supported by grants from The Bank of Sweden Tercentenary Foundation, The Swedish Council for Research in the Humanities and Social Sciences and The Swedish Board for Technical Development.

REFERENCES

TOWARDS AN ACCOUNT OF LANGUAGE-SPECIFIC PATTERNS OF THE TIMING OF VOICING

C. Wills and G. Docherty

Department of Speech, University of Newcastle-upon-Tyne, UK.

ABSTRACT

Results are presented from a study of the timing of voicing in English obstruents produced by native speakers of French and Spanish. It is suggested that in attempting to account for the timing of voicing (in native as well as non-native performance) an incomplete picture may be obtained if the variability in speaker performance is omitted from the account, and if excessive attention is focused on VOT as opposed to overall laryngeal-supralaryngeal coordination.

1. INTRODUCTION

Instrumental studies of the phonetic performance of non-native speakers of a language (particularly English) have been used (as evidence) in support of a model of acquisition of L2 (e.g.[2]), and (b) to shed light on the status of fine-grained phonetic variability, specifically regarding the extent to which it is a language-specific and learned aspect of phonetic performance (e.g. [3,6]). Studies of consequent production in L2 have focused almost exclusively on VOT, and the basis of comparison between different groups of subjects has typically been the mean VOT for particular categories of stops.

In this communication it is suggested that by commonly adopting the approach just described, previous studies may be overlooking some of the fundamental characteristics of the L2 (and L1) speaker performance. It has recently been proposed that the phonetic representation of an utterance may consist not of a string of precise target specifications, but may instead be characterised by built-in variability and underspecification [1,4]. The implication of this is that an account of performance focusing exclusively on mean scores may only be painting part of the picture. Furthermore, recent work on the detailed characteristics of laryngeal timing in stops [5] and on the phonetic and phonological representation of the voicing contrast [1] suggests that greater observational and explanatory emphasis should be placed on the overall timing and coordination of laryngeal and supralaryngeal gestures, and that variability of VOT (for example) may arise from variability of other timing and control parameters as opposed to being directly manipulated itself.

It seems timely therefore to investigate whether these revised notions of target and control would lead to the timing of voicing lead to rather different inferences being made about speech production from data obtained from L2 performance. This is the aim of a project being undertaken at the University of Newcastle-upon-Tyne, and the goal of this paper is to present a snapshot of some early results.

2. PROCEDURE

The aim of the experiment described below is to study the production of /p/, /b/, /v/ and /z/ in English by native speakers of French and Spanish. This paper deals only with the results pertaining to /b/ and /z/. Five native speakers of French and three of Spanish were recruited. The French speakers all lived in the North-East of England for over 8 years. Due to difficulties in locating subjects, the group of Spanish speakers was rather heterogeneous (a factor to be borne in mind in interpreting the results). SP1 had lived in the UK for 2 years, SP2 for 20 years, SP3 for 15 years. A group of 5 native English speakers was used for control purposes. Henceforth the subjects are referred to as ENG1-5, FR1-5 and SP1-3.

All the speakers were recorded producing (a) a list of 16 isolated English words (5 repetitions) (b) at least 4 cases each of initial /p,b,v,z/; (b) the same words embedded in a carrier sentence (5 repetitions). The FR and SP speakers were also recorded producing 5 repetitions of a matched set of isolated (16) French and (12) Spanish words respectively (the Spanish list was shorter due to absence of initial /z/ in Spanish) and the same words embedded in a French or Spanish carrier sentence (only 3 repetitions of the carrier sentences were obtained from SP1-3). The conditions are referred to henceforth as (1) Eng/Eng (i.e. English subjects/English words or sentences) (2) Fr/Eng (3) Fr/Eng (4) Sp/Sp (5) Sp/Eng. High quality DAT recordings were made in studio conditions. Subjects were asked to read the material from printed lists at a comfortable rate.

Wide-band spectrometers were made of the data using a LSI Speech Workstation, and were displayed on the screen of a PC terminal aligned with the corresponding speech waveform. The following measurements were taken for each token: VOT (stops only) taken as the interval between the release burst of a stop and the onset of the first vertical vibration for a following vowel; voice onset duration defined as the interval between the release burst of the stop and the point at which the second and higher formants disappeared from the spectrogram during the transition from the preceding vowel (this measurement could only be performed in the carrier sentence conditions given the need for a preceding vowel context); fricative duration defined as the interval between the onset and offset of the noise component visible in the spectrogram corresponding to the fricative; medial voicing, defined as the presence of vertical vibrations during the intervals previously identified as a stop or fricative.

3. RESULTS

Space prevents a detailed exposition of the results. Table 1 shows the mean VOT, consonant duration and medial voicing for /b/ and /z/ produced under the various conditions by the FR and SP speakers only. The principal findings are as follows.

- In /b/ in isolated words both negative and positive VOTs are found in Fr/Eng, Fr/Eng and Eng/Eng. Eng/Eng stops have few negative VOTs, Fr/Eng rather more, and Fr/Eng the most. The results for Sp/Eng speakers differ according to the subject. SP1 produced only prevocalic (i.e. negative VOT) stops in both languages, SP2 produces both short lag VOT and prevocalic stops in both languages, but with a difference in weighting such that the VOTs are predominantly short lag in the English words and negative in the Spanish words. SP3 uses both patterns in English without any apparent weighting, but produced almost exclusively prevocalic stops in Spanish.

- In /z/ in carrier sentences, both the English and French subjects' performance is characterised by a good deal of variability. On the whole, the Fr/Eng stops are more 'voiced' (i.e. more commonly entirely voiced, and with generally proportionally longer intervals of medial voicing) than stops produced in either the Eng/Eng or Fr/Eng conditions. The stops in the latter two conditions are similar with the exception that the Fr/Eng stops are considerably longer on average. There are large differences in the realisation of /b/ by SP subjects across the two languages. SP1 and SP3 in the Sp/Sp condition produce /b/ predominantly as fully voiced or partially voiced tokens of /b/. SP2 produces both fully voiced and partially voiced stops across all three conditions.

- In /z/ in isolated words the principal feature is the variability in the data. French and English subjects produce predominantly fully voiced or partially
Table 1: Mean VOT, consonant duration (CD), medial voicing (MV), and no. of cases of (A) complete devoicing, (B) partial devoicing and (C) complete voicing observed in /v/ and /z/ by FR and SP speakers (figures in parentheses are standard deviations/number of cases). All means and s.d.s are given in ms. Shortfalls in n reflect cases where either speaker error or measurement uncertainty forced exclusion of a token.

<table>
<thead>
<tr>
<th>Mean VOT in /v/ in isolated words -- separate means given for vo and vo VOT</th>
<th>Fr/Fr</th>
<th>Fr/Eng</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sp1 19(4/0)</td>
<td>71(13/11)</td>
<td>216(17)</td>
</tr>
<tr>
<td>Sp2 17(3/0)</td>
<td>94(14/7)</td>
<td>211(27)</td>
</tr>
<tr>
<td>Sp3 24(15/0)</td>
<td>44(24/0)</td>
<td>164(14/0)</td>
</tr>
<tr>
<td>Sp4 17(1/1)</td>
<td>80(22/15)</td>
<td>127(4/2)</td>
</tr>
<tr>
<td>Sp5 12(18/1)</td>
<td>62(22/9)</td>
<td>134(12/6)</td>
</tr>
</tbody>
</table>

Mean consonant duration, medial voicing and summary of timing patterns for /v/ in carrier sentences

<table>
<thead>
<tr>
<th>Mean consonant duration, medial voicing and summary of timing patterns for /v/ in isolated words</th>
<th>Fr/Eng</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Sp1 74(14/20)</td>
<td>74(14/20)</td>
</tr>
<tr>
<td>Sp2 91(36/20)</td>
<td>90(18/20)</td>
</tr>
<tr>
<td>Sp3 67(7/20)</td>
<td>61(14/20)</td>
</tr>
<tr>
<td>Sp4 103(22/20)</td>
<td>100(24/20)</td>
</tr>
<tr>
<td>Sp5 9(21/19)</td>
<td>71(19/19)</td>
</tr>
<tr>
<td>Sp/Sp</td>
<td>Sp/Eng</td>
</tr>
</tbody>
</table>

5 REFERENCES


Devoiced /z/ in all three conditions. No trends emerge regarding changes produced by French speakers in their performance of Fr/Eng. Two features emerge from the Spanish data (in which, of course, speakers face a novel situation given the absence of word-initial /z/ in Spanish). SP1 consistently initiates phonation before the start of the fricative noise characterising the /z/ (mean voicing lead = 121ms -- not shown in Table 1), and on some occasions proceeds to produce a fricative without any phonation, whilst on others voicing continues all the way through the fricative into the following vowel. In all the SP subjects there is a tendency for there to be a larger number of cases of completely devoiced /z/ in the Sp/Eng condition than in the Eng/Eng condition.

In /z/ in the carrier sentence condition, variability in realisation is the principal feature, with cases of full voicing and partial devoicing found across all the tokens, and with full devoicing being found somewhat less frequently. In the Fr/Eng condition, there is a tendency for /z/ to have shorter intervals of medial voicing than are found in Fr/Eng tokens. Completely voiceless tokens of /z/ are found more commonly in Sp/Eng than in Eng/Eng.

4. DISCUSSION

The absence of data from monolingual French and Spanish speakers precludes at this stage a statement regarding the degree of interaction of L1/L2 in the data (e.g. along the lines described in [2]), but the results do confirm the findings of amongst others [3] and [6] that the fine detail of phonetic realisation may be altered in the production of consonans in L2. The results conform to previous studies showing that VOT is one parameter which can be observed to alter in L2 performance. The data pertaining to /z/ shows that speakers are also able to manipulate laryngeal-supralaryngeal timing in the production of other sounds. In the light of this, we would seem fruitful to work towards a broader account of this aspect of non-native speaker performance than has been offered so far, recognising that VOT is a reflection of a more general process of gesture coordination, and thereby approximate an account which covers timing of voicing in general as opposed to only in stops.

The data also suggests that an account of the speakers' performance which presented no more than the mean VOT, consonant duration, and medial voicing would only paint part of the picture, and in particular would obscure the abundant inter- and intra-subject variability observed in the data, and consequently one of the major features of that data. For example, the mean medial voicing for Fr1's /z/ in Fr/Eng isolated words would not be a good reflection of the fact that half of the tokens produced by Fr1 are completely devoiced, and almost all the remainder are fully voiced. The observations made in this study could only be fully characterised by consideration of some measure of variance. Observations expressed in this way will allow full evaluation of the subjects' performance in the light of the work mentioned in 1 regarding inter-sentence variability in phonetic targeting. This work is now in progress.
ABSTRACT
Phoneticians can now make much of their laboratory apparatus into the field. Tape recorders have long been available, but their utility is much increased when they are used in conjunction with a portable computer. The computer not only provides convenient editing and play back facilities, but also can produce spectrograms, pitch curves, and other acoustic analyses. In addition, physiological parameters such as pressure and air flow and electromyographic data can be recorded and analyzed in the field on a portable computer. Photography (including video recording) and palatography are further tools for field use.

1. INTRODUCTION
There is a story about Daniel Jones, the great British phonetician who dominated the field in the first half of this century. When he was about to go off on a field trip someone asked him what instruments he was going to take with him. He pointed to his ears and said: “Only these.” It is surely true that by far the most valuable assets a phonetician can have are a trained set of ears. It is also true (and Daniel Jones would certainly agree) that the ears should be coupled to highly trained vocal organs that are capable of producing a wide range of sounds. There is no substitute for the ability to hear small distinctions in sounds. There is also no substitute for the ability to pronounce alternative possibilities, so that one can ask a speaker which of two pronunciations sounds better. One of the most efficient procedures for getting results in the field is to test different hypotheses by trying out various vocal gestures of one’s own. Nevertheless, however well trained they might be, phoneticians who now go out with only their ears and their own vocal apparatus are doing themselves a disservice.

2. RECORDING
What sort of machine should be used for making field recordings? As portable computers become more available, the days of dependence on tape recorders may be passing. Direct recording onto portable computers may be used, with the tape recorder being regarded simply as a backup. The computer system should be capable of sampling speech at 20,000 Hz for high quality listening and analysis, and at 10,000 Hz for the analysis of vowels and similar sounds. Even when considered just as devices for reproducing sounds, computers are much more versatile than tape recorders. Fieldworkers want to be able to record word lists or short paragraphs and then to play back selected pieces over and over again, so that they can hear subtle nuances of sounds that are new to them. They also want to be able to hear one sound, and then, immediately afterwards, hear another that may contrast with it. Both tasks can be done somewhat cumbersonely and tediously using tape recorders. But they are trivial, normal operations on any computer equipped with a means for digitizing and editing recorded sounds.

3. ACoustIC ANALYSES
In addition to being useful as a sophisticated playback device, a computer can provide several types of analysis that a fieldworker might find useful. The best display of the general acoustic characteristics of a sound is a spectrogram. Figure 1 shows the kind of spectrogram that can be produced on a portable computer without a color (gray scale) screen, printed on a light weight battery operated printer used in the field. The display in Figure 1 was created by a commercially available program, Signalyze. This program should not be judged by the spectrogram in this figure; the spectrograms it can generate on a color screen on a laboratory computer are much more impressive. But even the display that it is possible to print in the field can be very useful. The words shown illustrate the four contrastive sibilants that occur in Toda, a Dravidian language spoken in the Nilgiri Hills in India. Each of these words ends in a different sibilant. The overall spectral characteristics of these sibilants are evident. The laminal dental sibilant at the end of the first word has the highest frequency, and the retroflex sibilant at the end of the last word has the lowest. The apical alveolar and (laminal) palato-alveolar sibilants at the ends of the second and third words have very similar spectral characteristics. (The lowering of the spectral energy peak at the end of the second word is a non-distinctive feature, being simply due to the closure of the lip for the consonant at the beginning of the next word.) These two sibilants are distinguished primarily by their on-glides. The increasing second formant at the end of the third word is due to the raising of the blade and front of the tongue for this retroflex sibilant. A great deal of information can be obtained even from these low quality spectrograms, produced under field conditions. Of course, still more information can be obtained from high quality spectrograms produced by this or another program on a laboratory computer at a later date. Another kind of analysis that is very useful to the fieldworker is one that indicates the pitch. The Signalyze program discussed above will also generate good displays of the fundamental frequency (and it will produce narrow band spectrograms, which are sometimes even more useful for pitch analysis when a creaky voice quality or other unusual spectral characteristics are involved). But a number of other programs will also provide similar information. Figure 2 shows the fundamental frequency in a set of words with contrasting tones in Sukuma, as analyzed by a public domain modification of SoundWave, written at the University of Uppsala, Sweden.
The final kind of computer analysis of speech sounds that will be discussed here is one for determining the formant frequencies, the principal aspects of vowel quality. A common way of obtaining formant frequencies is by inspection and peak picking using superimposed LPC and FFT displays. The Uppsala software mentioned above provides a convenient way of producing displays of this kind in the field. When making an FFT it is important to remember the system limitations. In effect, an FFT provides the amplitudes of the spectral components that are present on the assumption that these components are all multiples of a wave with frequency depending on the number of points in the FFT. The greater the number of points in the FFT, the longer the wave length, thus the lower the frequency of this wave, and the smaller the interval between calculated components. But any program calculating an FFT will have a certain maximum number of points permissible (usually something like 512 or 256). Accordingly, the only way to further increase the accuracy in the frequency domain (i.e., to decrease the interval between measured components) is to decrease the sample rate. This will have the effect of decreasing the range of frequencies that can be observed. But it will also mean that all the components calculated will be within that range. Given a 512 point FFT and a sample rate of 20,000 Hz, there will be 256 components spaced about 40 Hz apart in the range up to 10,000 Hz. But if the sample rate is reduced to 10,000 Hz, the components in the same FFT will be spaced about 20 Hz apart in the range up to 5,000 Hz. It was for this reason that it was suggested earlier that if vowel formants were being studied it is advisable to use a lower sampling rate. The alternative would be to use an FFT with a larger number of points, but no analysis system will permit the maximum number of points to be increased beyond some fixed limit.

4. PHYSIOLOGICAL DATA

Acoustic analyses made from good quality tape recordings can provide large amounts of data. But they often do not indicate in an unambiguous way important articulatory facts such as the degree of nasalization, the phonation type, the direction of the airstream or the timing of movements of the vocal organs. The best way of gaining information on these phonetic parameters is by recording a number of aerodynamic parameters, using a portable computer. The general form of the system we use is shown in Figure 3. We can record the audio signal and up to three physiological signals. Typically these include one pressure (either the pressure of the air in the pharynx obtained by passing a tube through the nose, or the pressure of the air in the mouth using a more convenient tube between the lips), and the oral and nasal air flow. This system provides good data on degrees of nasalization. We have also used it to record an approximation to the subglottal pressure by means of a tube with a small balloon on the end of it in the esophagus, in investigations of prosodic features. Electrogallographic data can be recorded in a similar way.

Fieldworkers want to know not only the manner but also the place of articulation. Photographs of the lips can be very informative particularly if a mirror is used so that a full face and side views are recorded simultaneously. Palatography is also a well known method of obtaining articulatory data that can be used in the field. The comparative simplicity of this technique should not disguise the fact that it is still one of the most useful ways of obtaining information on the place of articulation and on distinctions between apical and laminal gestures. A useful way of recording the (static) palatographic records is by means of a video camera, which can also be used for recording the (dynamic) movements of the lips as mentioned above. Video images can easily be transferred to a computer, where they can be analyzed and measured — all while still in the field.

Finally, it should always be remembered that Daniel Jones was right. All the paraphernalia of the modern phonetics laboratory can never replace the human observer.

My thanks are due to Tony Traill for his wonderful collaboration in an earlier version of this paper.
AN ACOUSTIC STUDY OF XHOSA CLICKS

Bonny Sands

Linguistics Department, UCLA, Los Angeles, USA.

ABSTRACT

Clicks in Xhosa, a Bantu language spoken in South Africa, are made with one of three front closures, and with one of five accompaniments. The dental and lateral click types are characterized by an affricated release, while the alveolopalatal click type is not. Coarticulatory relations between clicks and vowels are less extensive than those between other consonants and their following vowels. Neither the front nor the back click closure varies much according to vowel context. The only coarticulatory effects seen are due to lip-rounding, which uses an articulator which is not involved in the production of clicks in Xhosa.

1. INTRODUCTION

There is much that is unknown about how clicks pattern with respect to other consonants. First, it is not clear whether clicks involve the same features as other consonants. And it is not clear whether the phonetic properties of these features are the same for clicks as they are for pulmonic consonants. An invariant acoustic property which is argued to exist for some feature or place of articulation should also exist for clicks sharing that feature or place of articulation. A feature such as [coronal] should have the same definition for pulmonic stops and fricatives and clicks. Unfortunately, the work on acoustic invariance [4] has largely ignored clicks in the determination of acoustic properties of features. Second, the way clicks interact with neighboring segments may be different from the way pulmonic consonants behave. Do clicks coarticulate with neighboring vowels?

2. CHARACTERISTICS OF THE FRONT CLICK CLOSURE

The data analyzed in this study were taken from a recording, kindly supplied by Professors Louw and Finlayson, of four male and four female Xhosa speakers saying words containing each of the 15 phonemic clicks before each of the vowels /i, e, a, o, u/ and /u/. Temporal characteristics of the clicks were also analyzed and are reported in [5]. The spectra in this study were made using a 25 ms window starting at the release of the consonant. Spectra were made on the DSP Sonograph using speech sampled at 40,960 Hz. Frequencies range up to 16,000 Hz. The power spectra of the click bursts of eight speakers for the voiceless aspirated, voiceless unaspirated and breathy voiced clicks before each of the five vowels were analyzed, giving 120 tokens of each click type. As the back click closure is released shortly after the release of the front closure, some noise from the back release may be included in the 25 ms window used.

The degree of coarticulation between a stop consonant and following vowel can be examined by comparing the spectral pattern of the consonant burst before different vowels. If vowel position is anticipated in the consonant, the burst will show modifications that echo some characteristics of the vowels.

2.1 SPECTRAL ANALYSIS

As seen in Figures 1 and 2, the dental clicks have a diffuse spectrum, and a great deal of energy above 6000 Hz. Dental clicks typically have energy present from 0 to 9000 Hz, and energy of having a diffuse spectrum, as would be predicted by [1,6]. As Figure 2 shows, tokens preceding the rounded vowels show a concentration of energy in the lower spectral region resulting from attenuation of amplitudes in the higher frequency range. The energy in the lower frequency band is greater in amplitude relative to the energy above 10,000 Hz for the clicks before rounded vowels. In particular, they show a peak of energy around 3000-4000 Hz.

For the alveolopalatal clicks, as seen in Figures 1 and 2, there is typically one main band of energy in the low frequency range, between 1000 and 1700 Hz. The frequency range of this band tends to be higher for the female speakers than for the males. Alveolopalatal clicks are non-anterior and have a compact spectral shape. This is similar to pulmonic coronal consonants which are not anterior, which are usually

---

Figure 1: Mean spectra of the dental, alveolopalatal and lateral clicks before the vowels /i, e, a, o, u/ for two male Xhosa speakers. Each curve is the mean of six spectra.

Figure 2: Spectra of voiceless unaspirated dental clicks of one female speaker of Xhosa, before each of the five vowels.

Figure 3: Spectra of voiceless unaspirated alveolopalatal clicks of one female speaker of Xhosa, before each of the five vowels.
characterized as having a compact spectral shape [1].

The effect of a rounded vowel on a preceding click can be seen for the alveolo-palatal clicks in Figure 3. As for the dental clicks, those preceding the rounded vowels show a concentration of energy in the lower spectral region, that is, below 2000 Hz. Energy occurs in a narrower band for the clicks preceding rounded vowels. The majority of tokens before the unrounded vowels have fairly prominent energy between 3800 and 4800 Hz, but the majority before rounded vowels do not. It may be that all alveolo-palatal clicks have audible energy in this range which does not appear in spectra designed to show the prominent peaks, as it is of such low amplitude relative to the low frequency band of energy.

The lateral click bursts, as seen in Figures 1 and 4, have a diffuse spectrum in the frequency range of 0 to 5000 Hz. They often have energy up to 8000 Hz or beyond, but it is typically of lower amplitude relative to energy below 5000 Hz. The energy in the spectrum is greatest in three broad frequency ranges, which are lower for male speakers than for female speakers. The spectrum can be delineated into regions presumably because of zeros caused by side cavities to the lateral channel of airflow. The majority of tokens before the unrounded vowels have energy present in the first range, between 1000 and 2000 Hz. The second region ranges from 2100-4600 Hz for female speakers, and from 2000-2900 Hz for male speakers. The third region ranges from 4000-4800 Hz for female and from 3000-4500 Hz for male speakers. As seen in Figure 4, the peak of energy which occurs below 2000 Hz tends to be at a lower frequency for clicks preceding a rounded vowel. The lateral click bursts share certain acoustic characteristics with other laterals. Lateral clicks and lateral approximants typically have energy at 3000 Hz and above. While lateral approximants typically have energy around 1200 Hz, the lateral clicks typically have a prominence between 1000 and 2000 Hz.

There were no consistent differences between the power spectra of any of the three click types before the vowels /i, e, a/. In particular, no consistent effect of the high front vowel /i/ is seen in the vowel which commonly causes extensive coarticulation effects with other consonants. There are however notable differences between the power spectra of the clicks preceding /i, e, and a/ and those preceding the rounded vowels /i/, /e/, and /a/ which is an expected result of coarticulation of the rounding of these vowels. Before rounded vowels, clicks show a shift in energy to the lower frequency region.

3. CHARACTERISTICS OF THE BACK CLICK CLOSURE

It may be that transitions into a following vowel are affected by click type. We might expect some information about click type to be contained in the vowel onset transitions, as this is often considered to be the primary cue for place of articulation of pulmonic stops. Alternatively, vowels following clicks might be expected to have all onset transitions which are indicative of a dorsal consonant since the release of the back click closure follows the release of the front one.

Measurements were made of formant transitions and vowel formants for the first three formants of the vowels /i, e, a, u/ occurring after dorsal, lateral, and alveolo-palatal voiceless unspirated clicks. The vowels of 7 Xhosa speakers were made. Formants were measured using LPC analysis on the Macintosh using UCLA/Uppsala Soundwave. A 256 point analysis window was used, and speech was sampled at 11 KHz. Formants were measured in the middle of the vowel and at the onset of voicing, and averaged.

![Figure 4: Spectra of voiceless unaspirated lateral clicks of one female speaker of Xhosa, before each of the five vowels.](image)

Figure 4: Spectra of voiceless unaspirated lateral clicks of one female speaker of Xhosa, before each of the five vowels.

No significant differences in the vowel formant onsets, were found for vowel by front click closure, using a 2-factor ANOVA. There is no significant acoustic evidence indicating that the vowel formant transitions vary due to type of front click closure. As seen in Figure 5, the difference between the onset for /i/ following each of the click types was very similar. The dentals show marginally lower F2 and F3 than the laterals, but these differences are not significant.

4. SUMMARY

Clicks have similar spectral characteristics to non-click consonants. Coarticulatory relations between clicks and vowels are less extensive than those between other consonants and their following vowels. However, this is not surprising, considering that the tongue body cannot freely vary its position in clicks because both the front and the back of the tongue have to be in particular positions to produce the consonant. Coarticulation involving the tongue position of vowels must be limited. This is similar to the constraints observed in vowel to vowel coarticulations across a consonant with a secondary palatal or velar articulation. The only coarticulation effect seen is that due to the anticipation of vowel rounding, since this does not involve a gesture used in the click production. These facts seem more compatible with a phonological theory in which the articulators are primary nodes [2] rather than features for place of articulation [3].

Many thanks go to Ian Maddieson in particular, and also to Pat Keating, Peter Ladefoged, Keith Johnson and John Choi for their helpful insights and comments.

5. REFERENCES

THE EFFECT OF LINGUISTIC EXPECTANCY ON PHONETIC TRANSCRIPTION:
DEVELOPING AN ADEQUATE ALIGNMENT ALGORITHM

C. Cucchiarii* & R. van Bezooijen**
* Department of Language and Speech, Nijmegen, The Netherlands
** Institute of Phonetic Sciences, Amsterdam, The Netherlands

ABSTRACT
This paper describes an alignment algorithm developed for transcription comparison. Theoretical and practical problems connected with the use of such a program are considered (1).

1. INTRODUCTION
A segmental transcription is the auditory analysis of an utterance into discrete units of sound represented by phonetic symbols. Such an analysis may be undertaken either to give a very detailed description of an utterance (allophonic transcription) or to indicate the distinctive categories of a language (phonemic transcription). Implicit in this distinction is the notion that the transcription is made by a transcriber who is familiar with the language to be transcribed. A different type of transcription may be obtained when a transcriber is required to transcribe an unknown language. The result is a so-called impressionistic transcription. The term impressionistic here refers to the fact that the transcriber has no recourse to the phonological system of the language being transcribed. All types of transcription, i.e. allophonic, phonemic, and impressionistic, have long been used in many fields of linguistic research as a means of recording speech material. However, the validity of these procedures has hardly ever been questioned. This is surprising, especially if we consider that analyses of this type are subject to the influence of a great number of variables relating both to the transcriber (experience, degree of familiarity with the language being transcribed, etc.) and to the type of speech under investigation (speech style, length of the utterance, rate of speech etc.).

In the light of these considerations we thought it would be useful to determine to what extent transcription performance can vary as a function of some of the factors mentioned above. Three variables were selected for investigation: 1. the transcriber's degree of familiarity with the language transcribed, 2. the presence of linguistic context, and 3. speech style. What these three variables have in common is that they are all related to linguistic expectancy, albeit to different degrees.

In the following section we will describe the method used, paying particular attention to the alignment program developed for transcription comparison and to the problems associated with the use of such a program. In section 3 preliminary results of its application will be presented.

2. METHOD
2.1. Transcription alignment
In order to determine the effect of the above-mentioned factors on transcription performance we need to be able to measure the difference between two transcriptions of the same utterance. Since phonetic transcriptions are linear sequences of symbols, the overall difference between two transcriptions of the same utterance is here defined as the sum of the differences between corresponding elements, i.e. symbols describing the same articulatory event. This implies that before two strings of symbols can be compared they have to be aligned, i.e. each symbol in one string has to be matched with the corresponding symbol in the other string.

Consider the enormous amount of material in our investigation (8640 transcriptions to be compared thousands of times) it was unthinkably to align transcriptions by hand. A program was therefore developed which makes it possible to align different transcriptions of the same utterance. The algorithm employed in our alignment program very much resembles the one developed by Picone et al. (2). This is an adapted version of the standard dynamic programming algorithm, which aligns two strings of symbols minimizing the sum of the insertion, deletion, and substitution costs.

The algorithm employed in our alignment program very much resembles the one developed by Picone et al. (2). This is an adapted version of the standard dynamic programming algorithm, which aligns two strings of symbols minimizing the sum of the insertion, deletion, and substitution costs. The algorithm is implemented in C and is available from the authors. The program can be used to align transcriptions of different languages, provided that suitable symbols are used for each language. The program is designed to align transcriptions of a fixed length, but it can be modified to align transcriptions of different lengths.

3. Apart from a few exceptions, both programs disallow vowel-to-consonant matches. In Picone et al. this is achieved by adding an extra matrix in which distances between vowels and consonants are greater than distances to the null symbol. A restricted number of matches between vowels and consonants is allowed by defining their distance to be lower than the distance to the null symbol. In our program vowel-to-consonant matches are prevented by rule. Possible exceptions are to be included in a separate list with their respective costs.

At best, an alignment program will perform as well as a human expert (1). Of course human performance does not mean a hundred percent correctness, as there can be string pairs which are simply difficult to align, even for an experienced phonetician. This may be the case when two transcriptions are very different, both quantitatively (number of symbols contained) and qualitatively (nature of the phonetic symbols).

When phoneticians align transcriptions by hand they use their knowledge of speech production and perception to arrive at what they think is the best alignment. Alternatively, when an automatic system is used its knowledge has to be externalized in the form of rules, constraints or costs, which tell the alignment program what to do. It is evident that even the best combination of rules and distance values cannot guarantee the performance level of a human expert, as the latter has access to much more information, can use his intuition and can be more flexible. In other words, we have to settle for something which can only approach human performance. This means that in any alignment program human corrections will eventually be required.

When an alignment program produces unsatisfactory output there are two possible solutions: 1. one can alter the output or 2. one can change the structure of the program (rules and distance values). Although the first solution would be the easiest, it is extremely ad hoc. Moreover, it may be argued that if the program produces an unsatisfactory output, the user can rely on the basis of the knowledge built in it. So, instead of manipulating the outcome...
then it is clear that the distance value between /θ/ and /m/ is too small in relation to that between /t/ and /m/. Also changing the distance values has its drawbacks. Theoretically, it is not correct since distance values are based on feature counting and therefore have their own motivation. From a more practical point of view, there should be no objection to using the outcome of the alignment program in order to improve the distance matrices, as we know them to be far from ideal.

With null symbols things are different. In this case, feature counting cannot be applied simply because null symbols have no features. As a consequence, the distance value between a phonetic symbol and a null symbol can only be motivated by the efficiency of the alignment program: as long as the alignment is correct, the null symbol values are also correct.

In the following section we will present some results of the application of our alignment program.

3. ADEQUACY OF THE ALIGNMENT PROGRAM: PRELIMINARY RESULTS

So far, the alignment program described above has been tested on 1680 transcription pairs. These were transcriptions made by fourteen Language and Speech Pathology students at the University of Nijmegen, in two experimental rounds. The material transcribed in the first round consisted of 120 speech fragments containing sequences of sounds across word boundaries, extracted from their original contexts so that they sounded like nonsense syllables. The fragments differed with respect to language variety (Dutch, a Dutch dialect, and an unknown language, viz. Czech) and speech style (reading vs. spontaneous speech). The material transcribed in the second round consisted of the same fragments, this time presented in their original contexts (usually two or three words). The transcriptions were made in accordance with the pre 1989 version of the IPA. As mentioned above, null symbols constitute a problem because one simply does not know what they should be assigned. Initially, we gave null symbols maximum values, computed on the basis of the distances between phonetic symbols. So, for vowel deletion we obtained a value of 10 and for consonant deletion a value of 15. This choice turned out to be not very felicitous for two reasons, one theoretical, the other practical. First, it is not clear why deleting a consonant should have a higher value than deleting a vowel. Second, when used as input to the alignment program these values produced a few instances of distorted alignment, in that matching symbols with vowels led to a smaller cumulative distance than matching them with consonants. In a second trial we adopted the value 15 for both vowels and consonants. As the alignment program aims at minimizing the cumulative distance between two strings, giving null symbols such a high value may result in alignments with an insufficient number of null symbols. Conversely, lower values may lead to alignments with too many null symbols. In order to get a general idea of how our program works we checked all alignments obtained to determine whether they were correct. Cases of incorrect alignments were classified as follows:

1. incorrect alignment due to an insufficient number of null symbols
2. incorrect alignment due to the insertion of too many null symbols
3. incorrect alignment due to incorrect distance values between segments
4. difficulty in finding the right correspondence between segments
5. difficulty in finding the right correspondence between segments

Out of a total number of 1680 string pairs, 87 (5.17%) turned out to be incorrectly aligned. The distribution observed was the following:

<table>
<thead>
<tr>
<th>error type</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>cases</td>
<td>7</td>
<td>71</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>

As is clear from this table the number of incorrect alignments of the second type is disproportionately high. This has two main causes. The first, which accounts for 52 cases, is the impossibility of matches between vowels and consonants. We expected this to be a problem and had already planned to use a list of exceptions (see section 2.1). First, however, we wanted to get an idea of the incidence of these cases. Now the question is whether the exceptions should be included in the program, which could have undesirable results for other string pairs, or whether they should be applied afterwards. The second cause, which accounts for 19 cases, is the incorrect matching of diphthongs with long vowels. In its present form, the program aligns the long vowel with the first part of the diphthong and then matches the second part with a null symbol. Since this appears counterintuitive, we will have to be charged by making it possible to match the whole of the diphthong with the long vowel. Apart from these cases, for which a solution has already been suggested, the number of incorrect alignments is small (0.95%). This would seem to indicate that the improvements proposed above, the program should work satisfactorily.

At this point another crucial question arises: are the distance values used for transcription alignment to be used also as an indication of error gravity? This question particularly concerns the values attributed to null symbols. For instance, in the case at hand the extremely high cost associated with null symbols led to satisfactory alignments, but it also had the effect of strongly influencing the average distance between transcriptions computed by the alignment program (for vowels and consonants separately). In fact, the transcriptions pairs with the highest dissimilarity scores were those in which null symbols had been inserted. In order to gain more insight into the effect of the null symbol value on transcription alignment, we let the program align same transcriptions again, but this time with an average value for null symbols, viz. 7. This led to exactly the same distribution as that presented in Table 1. Obviously, the value 7 is to be preferred to 15 because it has less impact on the distance measure and still produces a high proportion of correct alignments. Even this lower value, however, has the effect of penalizing null symbol insertion. Of course this need not be wrong. If one thinks that omission segments or inserting them is a serious mistake then it is right to associate a high cost with null symbol insertion. Perhaps one would like to introduce gradations in the cost of right and left, so that omitting certain segments is considered more serious than omitting others. In general, one cannot a priori exclude the possibility that under certain circumstances it may be appropriate to adopt different values for transcription alignment in the transcription evaluation. Each case will have to be considered separately and the outcome will depend on the purpose of the transcription.

4. REFERENCES


(1) This research was supported by the Foundation for Linguistic Research, which is funded by the Netherlands organization for research, NWO.
PHONETIC TRANSCRIPTION AS A MEANS OF DIAGNOSTICALLY EVALUATING SYNTHETIC SPEECH

R. van Bezaaijen* and W.H. Vieregge**
* Institute of Phonetic Sciences, Amsterdam, The Netherlands
** Department of Language and Speech, Nijmegen, The Netherlands

ABSTRACT

This paper explores the possibilities of using narrow transcriptions as an enriched alternative to an open response identification test in the evaluation of synthetic speech at the segmental level. To that end, transcriptions of synthesized phonemes were compared with the corresponding identification data. It is concluded that transcription should not be used in place of but rather in combination with an identification test.

1. INTRODUCTION

Probably the best known test for evaluating synthetic speech at the segmental level is the Modified Rhyme Test (MRT) [6], used extensively for the comparative evaluation of American English synthetic speech systems. In the MRT, initial and final consonants are tested separately with meaningful English CVC words. For each stimulus word the listeners are presented with six alternatives, from which they have to choose the correct response. Although the MRT has several advantages, such as speed and ease of administration to untrained subjects, it has been criticized in the literature, especially with respect to the restrictions imposed on the responses and the limited phonetic coverage in which the target consonants are presented [cf. 3]. The objections raised are particularly serious i.e. the test is to be used for diagnostic purposes, i.e. to assess the flaws of a system with a view of improvement, rather than comparative purposes, i.e. to relate a system's overall performance to that of other systems or other variants of the same system. An alternative approach, adopted regularly in the diagnostic evaluation of synthesis of European languages [e.g. 2,7] is to use an open response task with a large stimulus set comprising both meaningful and meaningless words of various structures, such as CVC, VCV, VCCV, and CVVC. In this way, the confusions found reflect true, unbiased perceptual characteristics of the stimulus sounds and information is gained on the intelligibility of phonemes in a wide variety of phonetic contexts. With the right equipment, the responses can be analyzed (semi-)automatically and presented insightfully in terms of percentage correct phoneme identification and phoneme confusion matrices. The subjects need to be trained in the use of an unambiguous notation system, but the time investment can be relatively small if foreign language students are used. Although the approach described can certainly be considered to be an improvement over the MRT in diagnostic evaluation, one could speculate whether it would not be possible to have an even more finely tuned measuring instrument. For it is not difficult to point out some characteristics of open response identification tests which in their turn limit the type and detailedness of the information yielded. For example, if the subjects perceive more than the intended number of input phonemes, they are forced to make a choice. Also, responses are limited to the phoneme inventory of the language in question. Deviations from standard, natural phoneme realizations (e.g. undue aspiration, excessively abrupt voice onset, inadequate segmental duration) cannot be indicated. Moreover, voice quality features, such as creak or whisper, are left out of consideration. Nevertheless, it could be argued that these types of information can be relevant to improve the segmental quality of synthetic speech, especially with respect to acceptability (naturalness, pleasantness).

If one wants to go further than improving synthetic phoneme quality from a purely functional point of view, i.e. in terms of identification, as the intended phoneme, one may consider taking recourse to highly trained listeners who have an extensive symbolic inventory at their disposal to denote subtle and deviant sound characteristics, without any previous training. The possibilities of this approach were first explored by Van Gerwen and Vieregge [5], who used the narrow transcriptions made by one experienced ear-phonetician to improve the quality of a text-to-speech conversion system for Spanish. More than 200 words were transcribed twice, the first time to assess segmental imperfections, the second time to check the effects of alterations. The present study was designed to gain insight into the relative merits of narrow transcriptions and data yielded by an open response identification task as means of diagnostically evaluating the segmental quality of synthetic speech. The comparison took place within the framework of the Dutch SPIN-ASSP program (1985-1990), which was set up to improve text-to-speech conversion for Dutch. First, methodological details will be given. Next, results will be presented and discussed.

2. METHOD

2.1 Open response identification task

In April 1990 a segmental intelligibility test was conducted to evaluate the output of seven synthesis systems for Dutch. For each system, 180 CVC words and 100 VCCV words, phonotactically permissible combinations of Dutch phonemes, were presented in an open response identification task. Most words were meaningless, a few were meaningful. Each phoneme was presented in seven phonetic contexts (for further details, see [1]). Eleven advanced students of English from the University of Nijmegen served as subjects. All had some practical knowledge of phonetics, specifically applied to the pronunciation of English, but none had any experience in listening to synthetic speech. They were paid for their participation.

Each CVC and VCCV stimulus word was presented once, with an interstimulus interval of 4 sec. The responses were typed on terminal keyboards. All consonants and vowels had to be identified, using a specially developed simple but unambiguous notation system. The task was an open response task in the sense that any combination of phonemes could be responded with, provided the number of phoneme responses corresponded with the number of numbered phonemes in the stimulus word. At a later stage, the subjects' responses were analyzed (semi-)automatically in terms of percentages correct phoneme identification and phoneme confusions. The identification task proper was preceded by a short training of 30 minutes in which the notation to be used was explained and practiced. Furthermore, in the actual identification task, each subblock of CVC and VCCV stimuli was preceded by 10 practice stimuli of the corresponding type and synthesis system.

2.2 Transcription task

A large part of the stimulus material presented in the identification task was transcribed by 30 students of Speech and Language Pathology from the University of Nijmegen as part of a comprehensive course in segmental transcription of pathological speech. They worked in pairs, each of the 15 pairs yielding consensus transcriptions for 70 CVC and VCCV words, 10 for each synthesis system.

Since it would have been too time-consuming to examine the transcriptions of all phoneme realizations, it was clear a selection had to be made for the purpose of the present study. It was decided to consider the transcriptions of one realization of each of the seven CVC and VCCV phoneme positions for each of the seven synthesis systems, i.e. the realizations of 49 target phonemes. In view of the special relevance of a good diagnosis for poor phoneme realizations, in each case the phoneme which had yielded the lowest mean intelligibility score in the identification task was selected. The intelligibility scores for the target phonemes varied considerably (between 0% and 84% correct), as a function of phoneme category (vowel versus consonant) and phoneme position, and synthesis system.

138 139
On the average, each target phoneme occurred in 5.9 different words, amounting to a total of 291 phoneme realizations. The students' consensus transcriptions of these phoneme realizations were checked by the second author, an phonetician, who experienced both in the transcription of normal and pathological speech. A small part of the material (about 15%) was transcribed by him alone. The transcription system used was the one described in [4], i.e. the Extensions to the International Phonetic Alphabet for the transcription of atypical speech.

3. RESULTS AND DISCUSSION

The neatest way to establish the relative merits of an identification test and transcription as tools for improving synthetic speech would be a pretest-posttest design in which the effects of alterations based on the outcomes of the two methods were independently assessed and compared. It may be clear that this approach is practically unfeasible. Instead, we decided to use the results from the identification task as a reference for establishing the possible usefulness of transcription as an alternative means in diagnostic evaluation. After all, synthetic speech is primarily developed to allow man-machine communication in various applications. So, a first prerequisite of synthetic output is that it can be understood by "normal" human listeners, that the sounds produced are interpreted in terms of the intended phonemes. Any segmental diagnostic evaluation method should be capable of showing to what extent this basic condition is fulfilled. In other words, if transcription is to be considered as a valid diagnostic tool the data it yields should agree with the identification results obtained in a segmental intelligibility test.

Ideally, in addition to this basic information, narrow transcriptions should yield more. However, as was stated before, the usefulness of this extra information for diagnostic purposes can really only be assessed by formally testing the perceptual effects of the resulting alterations applied to the system in question. In the present study all transcription details throwing light on particular synthesis characteristics were considered as potentially useful on two conditions, (1) that they were systematic, i.e. occurred in at least 5% of the transcriptions pertaining to the realizations of a particular target phoneme, and (2) that they could not be inferred from the results yielded by the identification task.

With these definitions of what constitutes basic and extra information in mind, the transcriptions were carefully examined. To facilitate generalizations, each series of transcriptions pertaining to the realizations of the same target phoneme were assigned to one of the following three categories:

1. Equivalent to the identification method, i.e., leading to the same qualitative and quantitative interpretation in terms of correct and incorrect phonemes.
2. More informative, leading to the same qualitative and quantitative interpretation and, in addition, providing extra information as defined above.
3. Misleading, leading to a qualitatively or quantitatively different interpretation, suggesting an adaption, or an underestimation of phoneme intelligibility.

The distribution of the (series of) transcriptions for the 49 target phonemes was 30, 7, and 12 in categories 1, 2, and 3, respectively. So, in 30 cases (61%), spread over all 7 synthesis systems, the transcription and identification methods were found to be equivalent in the sense that they yielded the same basic information in terms of correct and incorrect phonemes. In 7 cases (14%), spread over 5 systems, transcription appeared to be more informative, providing extra information which was considered potentially useful for the improvement of the segmental quality of the synthetic system at hand. The information pertained to voice quality (3 cases), to the undue presence of a final consonant in VCVV words (2 cases), to diphthongization (1 case), and to overly strong phoneme realization (1 case).

In 12 cases (24%), spread over 6 systems, the transcriptions were misleading in the sense that they did not correspond with the pattern of responses obtained in the identification task. In 7 cases the difference was qualitative, in 5 cases quantitative. Of the latter, 2 would have led to an overestimation and 3 to an underestimation of the target phoneme intelligibility. We were somewhat amazed by the relatively high number of category 3 cases, since we had expected the transcriptions to generally show the same phoneme distribution as found in the identification task. The point was not clarified by an inspection of the original, untranscribed transcriptions, since the differences found hardly affected the categorization (there was only one doubtful case).

In any case, the outcome of the present study suggests that it is somewhat risky to use narrow transcriptions made by highly trained listeners as a substitute for an open response identification task with moderately trained listeners. Apparently, the transcriptions are not always a good predictor of the communicative adequacy of a system in terms of phoneme categorization. Moreover, the transcription approach has other disadvantages as well. For instance, it can be utilized by skilled listeners who have been trained extensively; the method is extremely time-consuming; the designer of the synthesis system has to be able to interpret the transcription symbols; and the data are very difficult to summarize in an insightful manner. This does not mean to say that we deny any role to transcription in the evaluation of synthetic speech. After all, the present study revealed several cases where transcriptions provided potentially useful diagnostic information not deducible from the results yielded by an open response identification test. The reader may recall that only those transcription details were categorized as potentially useful that occurred systematically in the transcriptions of the realizations of the same target phoneme. This is a rather strict condition, and it cannot be excluded that much more potentially useful information was contained in the transcriptions of individual items.

We are convinced that narrow transcription can contribute significantly to the improvement of synthetic speech, if it is used with specific questions in mind, i.e. at a more "local" level. One could think, for example, of a configuration in which a system developer consults one or more transcribers to test the validity of specific hypotheses based on his own perception - after all, it is a well-known fact that system developers generally lose objectivity when listening to the output of their own system - or, perhaps even better, to clarify the outcomes of a formal identification test. In our experience, the efficiency of this procedure is enhanced if the written transcriptions are accompanied by oral explanations.

REFERENCES


CONSONANT CLUSTERS: A COMPARISON BETWEEN WORD INTERNAL AND WORD JUNCTURE

Christine Meunier
Institut de Phonétique, Aix-en-Provence, France

ABSTRACT

We analyse the acoustic organisation of French consonant clusters (with two consonants) in three contexts: word internal position, word juncture provided with major boundary and word juncture provided with minor boundary. We use a specific classification of consonant clusters. Durations and acoustical transitions between both consonants are analysed in this paper.

1. INTRODUCTION

Some studies describe the acoustic and/or articulatory structure of the consonant structure[4][5]. The aim of our study is to evaluate the acoustic differences which can appear between a French word internal consonant cluster (two consonants) and the same cluster linking two words. We suppose that the acoustic features, we observed in word internal clusters, may support modifications if we change the boundary between the two consonants.

2. CONSONANTS AND CONSONANT CLUSTERS

We classified the French consonants in order to draw up a consonant cluster (GC) classification.

2.1. Consonant classes

- Stops: /p/ /b/ /t/ /d/ /g/ /k/ /t/ /d/ /g/ /k/
- Fricatives: /f/ /v/ /s/ /z/ /θ/ /ζ/ /j/ /r/
- Vocalic consonants: glides /j/ /w/ /l/, liquids /l/ /l/ and nasals /m/ /n/ /n/.

2.2. Consonant clusters classification

[1] We divided the GC into two groups: Homogeneous consonant clusters (both consonants belong to the same consonant class), and heterogeneous consonant clusters (both consonants belong to different consonant clusters). In these two groups, three types of GC can be deduced from the consonant classification:

Homogeneous GC:

- Ho1 (no fricatives + fricatives)
- Ho2 (no voicing + voice consonant)
- Ho3 (no voice consonant + voice consonant)

Heterogeneous GC:

- He1 (stop + fricatives)
- He2 (fricatives + voice consonant)
- He3 (voice consonant + fricatives)

3. SPEECH MATERIAL

We selected two corpora. In the first, the Word Corpus (CM, "Corpus Mors" in French), the GC are word internal; word initial for the heterogeneous groups (plz) and medial for the homogeneous ones (obus). We took into account only those GC from French lexical words. All the words are included in the sentence, "Ce n'est pas qu'il faut dire". In the second corpus, the Juncture Corpus (CJ, "Corpus Juncture" in French), we considered two levels of junctures: the first in a major boundary and the other in a minor boundary. In fact, the sentences of CJ follow the very simple syntactic structure: SN+SV. The first type of juncture (CJa) is between SN and SV (the major syntactic boundary), the second (CJb) is inside SV (between V and N, the minor boundary). We expected to obtain different acoustic effects with regard to the type of juncture which separate the first and the second consonant (C1 and C2). As a consequence, for each GC we analysed a triple coincidence: C2: "ce n'est pas qu'il faut dire" CJa: "équipe malentendu s'il vous plaît" CJb: "ce retard handicapé Robespierre"

We recorded two speakers (male) who read the three corpora twice. The total number of recorded words is 336 (112 for each corpus).

4. ACoustical analysis

4.1. Duration

We observed the variations in duration between CM, CJa and CJb (means and coefficient of variation) for the consonant clusters (duration of C1, C2 and CC) and for each consonant class. In the same way we compared the correlations of the durations of CC/CJa, CC/CJb, CJa/CJb, for each class of CC and for all together.

4.2. Transition phase

It is important point in the study of the consonant clusters is to observe the transition phase between C1 and C2. Two possibilities are considered:

The Direct Passage (PD): the GC is composed of C1 acoustical characteristics + C2 acoustical characteristics without any other segment.

The Transition Segment (ST): a segment different from the acoustic characteristics of C1 or C2, appears toward the boundary, it can be either a transformation or an insertion. In order to evaluate the distribution of the Transitory Segments, we have to draw up the acoustic characteristics of each consonant class:

- Stop: silence (or voicing with regard to the phonological description) and burst.
- Fricatives: noise with a stable specific frequency (voiced or unvoiced).
- Vocalic cons: voiced formant structure.

Any possible variations of these simple descriptions (with regard to the phonotactical transcription) will tell us if the transition phase is PD or ST realised.

5. HYPOTHESES

When we defined the Juncture Corpus we drew up hypotheses about the acoustical variations brought by the boundary degree between C1 and C2.

- the data of CJb would be closer to the data of CC (as long as we consider that the word boundaries disappear in continuous speech in French).
- the CJb clusters would be longer than the CJa ones (as long as the major boundary acoustic effect could be a duration increase of C1, C2 or both).

The disappearance of ST in the CJa clusters (as long as the ST presence is a cue for strong coarticulation), and the apparition of pauses between C1 and C2 (evidence of a major boundary).

The increase of partial and total assimilation numbers in the CJa clusters, and decrease in CJb ones (comparing them to CC clusters).

The results of the acoustical analysis will confirm or not our hypotheses.

6. RESULTS

6.1. Mean duration

Table 1: Mean duration (M) and coefficient of variation (C) of all the consonant clusters for the three corpora:

<table>
<thead>
<tr>
<th></th>
<th>CM</th>
<th>CJa</th>
<th>CJb</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>106</td>
<td>97</td>
<td>88</td>
</tr>
<tr>
<td>C2</td>
<td>122</td>
<td>127</td>
<td>129</td>
</tr>
<tr>
<td>CC</td>
<td>225</td>
<td>238</td>
<td>236</td>
</tr>
</tbody>
</table>

In the three contexts, C1 is always longer than C2, but the difference seems to decrease in the CJa context. The general means of CJa are slightly longer than those of CJb. We can explain the long durations of CC remaining that the CC clusters always belong to accented syllables.

Table 2: Mean duration (M) and coefficient of variation (C) of consonant classes in C1 position (C1, position (C2) and in general (STOP, PRI, VOC) for the three corpora:

<table>
<thead>
<tr>
<th></th>
<th>CM</th>
<th>CJa</th>
<th>CJb</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>104</td>
<td>97</td>
<td>88</td>
</tr>
<tr>
<td>C2</td>
<td>122</td>
<td>127</td>
<td>129</td>
</tr>
<tr>
<td>CC</td>
<td>225</td>
<td>238</td>
<td>236</td>
</tr>
</tbody>
</table>

We do not notice changes in the three corpora for stops: stops are always longer in C1 than in C2 position. For fricatives, we see a difference between CC and CJa (a and b) /CJa fricatives would be longer than in first than in second position; in CJa (a and b) tend to have the same duration whatever their position. Vocalic consonants are longer in first than in second position in CC, we notice the same
for Cfl (but with a smaller difference), and the opposite for Cj2. We must notice the strong stability of CC (whatever the consonant class), and the similarity between Cfl and Cj2 with the exception of vocalic consonants. Consonants seem to be longer in Cfl than in Cj2.

6.2. Correlations:
Table 3: Correlation matrix of C1, C2 and consonant classes for the three corpora in general (number: 92).

<table>
<thead>
<tr>
<th></th>
<th>CM</th>
<th>CjA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cfl</td>
<td>0.31</td>
<td>-0.18</td>
</tr>
<tr>
<td>CjB</td>
<td>0.36</td>
<td>0.256</td>
</tr>
<tr>
<td>CjC</td>
<td>0.39</td>
<td>0.217</td>
</tr>
</tbody>
</table>

Significant correlations for 0.01 and 0.02 probability: CM/Cfl, CjB/CjA (for C1), CM/CjB for (C1, C2 and GC), Cfl/CjB for (C1). Not significant: CM/CjA (for C2, GC).

Table 4: Idem table 3: H01 (number: 16).

<table>
<thead>
<tr>
<th></th>
<th>CM</th>
<th>CjA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cfl</td>
<td>0.152</td>
<td>-0.073</td>
</tr>
<tr>
<td>CjB</td>
<td>0.142</td>
<td>0.365</td>
</tr>
<tr>
<td>CjC</td>
<td>0.194</td>
<td>0.064</td>
</tr>
</tbody>
</table>

Significant correlations for 0.01 and 0.02 probability: CM/CjB (GC only). Not significant: CM/CjA, CM/CjB (for C1), CjB/CjA.

Table 5: Idem table 3: H02 (number: 20).

<table>
<thead>
<tr>
<th></th>
<th>CM</th>
<th>CjA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cfl</td>
<td>-0.028</td>
<td>-0.135</td>
</tr>
<tr>
<td>CjB</td>
<td>0.091</td>
<td>0.203</td>
</tr>
<tr>
<td>CjC</td>
<td>0.037</td>
<td>0.306</td>
</tr>
</tbody>
</table>

Significant correlations for 0.01 and 0.02 probability: none. Not significant: all.

6.3. Transition Phases
Table 8: Distribution of the Transitory Segments in the six consonant cluster classes for the three corpora; voicing opposition (opp de vat) and similar voicing (vat simi) inside clusters are separated in each class.

We observe a great proportion of ST when the two consonants are differently voiced; here the voiced consonant is in general partly (or more rarely, completely) devoiced. When the consonants are not in voicing opposition, some ST are also present: it can be an insertion of a vocalic element (particularly in H01), or the "consontanification" of the vocalic consonant (if/ following stops or fricatives). We did not note any pause in Cfl context.

7. Conclusion
Some of our hypotheses seem to be partially confirmed by the results of the acoustical analysis; Cfl clusters tend to be longer than CjB ones; the acoustic organisation of CjB clusters tends to look like CM one, instead of Cfl. In fact, acoustic organisation seems to be more stable when clusters are inside a word; but we must specify that the sentence in CC was always the same, it could also stabilise the GC production. Stability also characterises stops and fricatives instead of vocalic consonants which are acoustically more heterogeneous.

References
SPEAKING WHILE INTOXICATED: PHONETIC AND FORENSIC ASPECTS

Angelika Braun
Landeskriminalamt Nordrhein-Westfalen, Düsseldorf, Germany

ABSTRACT

Although there is a lot of everyday knowledge about the effect of alcohol on speech production, scientific studies on the subject are sparse. In the experiment reported here 33 subjects read a given text in sober condition and in intoxicated condition. The results show a marked increase in speech errors, a decrease in readiness to correct errors, as well as a number of segmental effects, e.g. lengthening and voicing changes in phonemes as well as forensic implications of the findings are discussed.

1. INTRODUCTION

It is common knowledge among phoneticians as well as phoneticians that the consumption of alcoholic beverages, especially in large quantities, affects the vocal tract in varying ways while the effect of alcohol on certain neurological mechanisms has been subject to a large number of investigations. Surprisingly little effort has been made among phoneticians and speech scientists to find out exactly what is the effect of alcohol on speech. One of the major shortcomings of the existing studies is that virtually all of them have attempted to measure the degree of intoxication. Instead, they often rely on the Widmark formula which only allows for a very rough estimate. Due to the difficulties in dealing with drunken subjects in an experimental situation, the number of subjects was usually very small, i.e. under 6 (e.g. 3, 5). Thus there is a number of very general findings indicating that speech produced under intoxication is slower, reduced in amplitude, and more error prone than speech produced in sober condition [3], but we are still in need of precise descriptions. The present study was motivated by this lack of data as well as the forensic application of phonetics, where the expert is often asked in court whether there is any indication of intoxication in a certain incriminating recording. One of the more unusual and spectacular cases in which the question of alcohol abuse was crucial concerned the Exxon Valdez oil spill. In cases like this it would not only be desirable to know exactly the effects of alcohol on speech production but also whether there is a correlation between the effects displayed and the amount of alcohol consumed. (This is of prime importance e.g. for the question of diminished responsibility).

2. EXPERIMENT

An experiment was carried out involving 33 main subjects who were 25 years old on the average (50 - 15 months). The task reported on here was the reading of a phonetically balanced text (The Northwest and the Sun) which was done in sober condition first. Subjects were then given 40% proof vodka. It was indicated to them that a blood alcohol concentration of between 0.1 and 0.2% was desirable for the purpose of the investigation and approximately how much vodka they would have to consume to achieve that. But there was no possibility to prescribe the exact amount they would have to drink. Thus, maximum alcohol levels of between 0.02% and 0.21% were actually achieved. The drinking time amounted to 90 minutes. 30 minutes later subjects were tested by means of a Siemens Alcomat breathalizer for their breath alcohol levels (which has a close to perfect correlation to blood alcohol level [1]) and subsequently read the text.

3. METHOD

A number of parameters including rate of articulation, fundamental frequency, segmental features and speech errors were investigated, the former by means of a computer program specially designed for speech analysis, the latter by auditory analysis. This presentation will, for reasons of time, focus on speech errors and selected segmental features.

4. RESULTS

4.1. Segmental features

There are some descriptions about the effects of alcohol on certain speech sounds, e.g. the amounts of /æ/ and //e// increase by auditory analysis. This presentation will, for reasons of time, focus on speech errors and selected segmental features.

4.2. Segment Lengthening

Segment lengthening forms one of the most consistently observed effects of alcohol. [2] The percentage of subjects showing vowel and consonant lengthening ranges from 15% (vowels as well as consonants)
Nasalization and denasalization

Figure 1

Number of incomplete articulations

Figure 2

Number of speech errors per speaker

Figure 3

below 0.06% to 50% (consonants) and 8.1% (vowels) above 0.08% max. BAL. Thus the steady-state portions of certain sounds seem to be increased at the expense of the articulatory precision of others.

4.2. Production Errors
Speech errors have long been used as an indicator for mental processing; therefore we also analyzed them in two different respects: (a) the number of speech errors (e.g. slippage of the tongue) in the real passage; (b) the readiness to correct the errors committed. There is a doubling of speech errors above a breath alcohol concentration of 0.08% and a drastic increase above 0.16% as compared to the sober condition. (Fig. 3) This means that even in a comparatively simple task like reading a text which does not involve cognitive planning, there is a significant increase at 0.08% alcohol level. The readiness to correct these errors which is commonly viewed as an indication of an internal monitoring mechanism was greatly impaired (i.e. reduced to about 1/3) even at very low levels of intoxication. There is no significant change up to 0.2%, but above that BAL, there are hardly any attempts to correct the errors at all. Also, there is a growing percentage of false corrections at high BALs, which amounts to over 30% of all corrections at BALs of 0.16% and above.

5. DISCUSSION
Alcohol is known to be neurotoxic, i.e. to impair coordination and nerve transmission. In speech, this results in a reduced and/or imprecise movement of two articulators which require the most precise control mechanisms: the tongue tip and the velum, whereas other sounds are sustained for a longer period than in sober conditions. With all of the parameters discussed here, the effect shows even at low BALs, but there is a marked increase above 0.08% and again at 0.16% (consonant denasalization; vowel length); or 0.20% (vowel nasalization; incomplete articulations). This seems to suggest that the effects of alcohol do not increase gradually but in steps. The study also shows that even in a reading task, there is a significant increase in the number of speech errors paralleled by a decrease in the attempt to correct the errors. This suggests that not only production processes are impaired but also the reception and comprehension of texts.

From the forensic perspective it has to be pointed out that even though most effects of alcohol are generally very consistent, there is always a small number of subjects who do not show them. Thus, there is no one-to-one relationship between the consumption of alcohol and the effects on speech in the sense that the presence of one (or better: several) of the impairments mentioned here point to an intoxication of the speaker but their absence may not be taken to prove sobriety.

6. REFERENCES
TEMPORAL CONTROL IN SPEECH OF CHILDREN AND ADULTS

Cecile T.L. Kuijpers

Institute of Phonetic Sciences, University of Amsterdam
Herengracht 338, 1016 CG Amsterdam, The Netherlands

ABSTRACT

Speech utterances of children and adults are compared with respect to phonologically short and long vowels, voiced and voiceless plosives, and the interaction between vowel duration and following consonant. It appears that four-year-old children and, to a lesser extent six-year-old children have not yet mastered the temporal control of these vocalic and suprasegmental segments. Results are interpreted in terms of a developing timing mechanism.

1. INTRODUCTION

From a segmental and suprasegmental point of view, acoustic-phonetic research of young children's speech utterances contributes to a better understanding of the development of speech motor control such as phonetic timing [1]. Phonetic timing concerns start and duration of phonetic intervals such as vowel duration, syllable duration, etc. [4]. One of the most appropriate instruments to investigate the timing mechanism in speech, as well as to study the development of this mechanism, are durational analyses of the utterances and their segmental constituents. Different linguistic factors will affect the duration of single phonetic intervals; concerning phonological features that serve to distinguish words (e.g., the short-long opposition in vowels, voiceing and also contractions of stress) length is one of the main characteristics and influences duration of phonetic intervals.

Concerning developmental research, several studies have shown that children have a slower speaking rate and that segmental durations are longer and variable than those of adults [7]. These temporal parameters approach the adult norm with increase in age [1, 5]. Most studies make use of an imitation procedure with nonsense words or a sentence repetition task. This in order to compare in a direct way young children's data to adult data and to control for the set of utterances across ages. However, the phonological features of the child's speech utterances will be reflected by durational values that are appropriate to his/her own developing mechanism [1]. Therefore, we chose to make use of spontaneous but controlled speech utterances instead of imitative speech. In this paper we want to emphasize two aspects that relate the linguistic parameters of 'vowel length' and 'voiceing' in Dutch to the phonetic-acoustic cues 'duration of the vowel' and 'duration of the closure'. As will be evident, short and long vowels differ in duration and closure whereas voice and voiceless plosives are characterized by long vs. closure duration.

Firstly, two basic questions can be formulated as follows: 1) how do young children handle durational values of short vs. long vowels as opposed to long and voiceless plosives and 2) how do they handle differences in closure duration of intervocalic voiced and voiceless plosives?

Secondly, the contextual effect of lengthening of the vowel preceding a voice consonant (short closure) and shortening of the vowel preceding a voiceless consonant (long closure) will be examined in the utterances of children and adults. This phenomenon, which is known as temporal compensation [4] is not inherent to the phonological system of Dutch but is considered to be an articulatory coordination. One of the claims to be made is that the temporal compensation between V and C is only mastered gradually by young children.

2. METHOD

2.1. Subjects

Four different age groups participated in the experiment: four-, six- and twelve-year-olds, plus adults. So far, only results of the two youngest age groups and adults are available and will be presented here. Each group consisted of six subjects, equally divided over male and female speakers. All of them were monolingual speakers of Dutch and none of them were judged to have any hearing loss or speech disorder. All subjects lived in the same area of the South-East of the Netherlands.

2.2. Material

Data are presented that refer to a set of 28 meaningful words. They are all two syllabic (C)VSCV(C) words with lexical stress upon the first syllable (+$syllable boundary). The intervocalic consonant was either a voiced plosive, that is /p/ or /t/, or a voiceless plosive, that is /f/ or /s/, e.g., words 'kabel' (cable) vs. 'spel' (pile). In approximately half of the words the vowel preceding the intervocalic consonant was a phonologically short vowel /a/ or /e/, /i/ or /I/, otherwise it was a long vowel /a/, /e/ or /i/. Experimental research with young children imposes several constraints upon the selection of meaningful words to be used: No minimal pairs could be found, 11 words of intervocalic voiceless plosives and 17 words with intervocalic voiced plosives were selected (among which optimally matched pairs), the initial consonants were not always identical and we had to make choices of one-morphemic as well as two-morphemic words to avoid an imitation procedure for all the words were elicited by picture cards.

2.3. Procedure

The elicitation procedure was based upon pictures drawn on separate cards. In all age groups we chose for the same procedure and all subjects pronounced the same set of words. The words were elicited by questions or sentences that had to be completed only by the word itself. This task would account for a spontaneous but controlled speech production without imitation whatever.

2.4. Recordings

Recordings of the four-year-old children were made at home with a Tandberg recorder and a microphone Sennheiser M211HN. The six- and twelve-year-old children and the adults were recorded in a laboratory setting with a Revox A77 recorder and an electroaryngograph to register the exact timing of the vocal pulsing. All recordings were recorded twice and both recordings were used for analysis. Even four- and six-year-old children pronounced 'correctly' 90% of both voiced and voiceless plosives, i.e., during segmentation both visual and auditory information indicated that neither substitution of voiced by voiceless plosives had taken place (and vice versa) for any deletion of intervocalic plosives.

2.5. Measurements

The synchronous audio- and electroaryngary signals were stored digitally on a microVAX II computer and the speech editing system provided visual and auditory information for verification. To be consistent in measurements we always concentrated upon the oscillographic signal using the traces of laryngeal activity for verification. In this paper we report on the following measures:

- vowel duration preceding intervocalic voiced and voiceless plosives
- closure duration and burst duration of the intervocalic plosives
- word duration

We do not want to dwell upon the criteria used for segmentation; they can be found in [2] and are in accordance with most criteria used in literature.

3. RESULTS

3.1. Vowel duration

Mean durations of the separate vowels, as well as mean durations of short vowels pooled and long vowels pooled, are presented in Table I. As can be deduced from the data, vowel durations between the age groups differ considerably.
Between the four- and six-year-old children no significant difference was found in overall vowel duration. Vowels of four- and six-year-olds were significantly longer than those of adults [F(1,10)=36.20; p<.001 and F(1,10)=35.42; p<.001]. Between four-year-olds and adults a 76% reduction of short vowels and a 47% reduction of long vowels was found; between six-year-olds and adults reductions of 52% and 36% respectively were found. The short-long opposition, which is an important phonological feature in Dutch, was clearly present in all age groups and the relative durational differences between short and long vowels was quite similar in the three age groups.

Table 1. Mean durations in ms of all vowels in the age groups. Below mean durations of short and long vowels are indicated as well as the ratio.

<table>
<thead>
<tr>
<th></th>
<th>Short vowels</th>
<th>Long vowels</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>147</td>
<td>239</td>
<td>0.63</td>
</tr>
<tr>
<td>6</td>
<td>125</td>
<td>217</td>
<td>0.59</td>
</tr>
<tr>
<td>Adults</td>
<td>81</td>
<td>159</td>
<td>0.54</td>
</tr>
</tbody>
</table>

In Fig.1 we have indicated this short-long opposition across ages. We can see that both types of vowels shorten in the same amount with age.

Regression analysis of the variable 'vowel duration' upon 'age' shows that the proportion of variation of short and long vowel duration can be perfectly predicted from age (R^2=.96 and R^2=.98).

3.2. Closure duration

Closure duration of the intervocalic voiceless plosives /p/ and /b/ are compared in Fig.2. As a measure of contrast, the ratio voiced/voiceless closure duration was calculated. In par. 3.1 we have shown that the ratio short/long vowel duration decreases with age, i.e. the contrast increases with age. Contrary to this, voiced closure is in consonant/Charts closure for /b/ vs. /p/ increases with age from 0.58 to 0.66 to 0.72 and for /v/ vs. /v/ from 0.50 to 0.59 to 0.68.

In Fig.2 we have indicated closure duration in ms for voiced and voiceless plosives in the three age groups. Each line represents vowel duration of one subject.

It will be clear that four-year-olds behave very differently from the older children and adults [F(1,10)=37.28; p<.001 and F(1,10)=36.20; p<.001], they do not make any distinction between vowel duration in a voiced or voiceless context. And, it is interesting to see that between the ages four and six a shortening of the vowel occurs only before voiceless consonants while vowel durations before voiceless consonants remain the same. Between six-year-olds and adults vowel duration reduces almost in the same amount whether preceding a voiced or a voiceless plosive. Analyses of covariance, with word duration being the covariate and a measure of speaking rate, indicated that differences could not be attributed to speaking rate alone. Probably, some effects due to age and to developmental structure also had an influence.

3.3. Vowel duration as a function of the following consonant

The three age groups were compared in their use of vowel duration as a function of the following voiced and voiceless plosives. In Fig.3a, b, and c behaviour of short and long vowels is plotted for all subjects in the three age groups.

4. DISCUSSION

Durational values of short and long vowels and voiced and voiceless closures in speech of three age groups were examined in relation to the phonological oppositions of 'vowel length' and 'voiceless'. The children's relative temporal structure of short vs. long vowel seems to be acquired before the age of four while relative closure durations of voiced vs. voiceless plosives is still in a developmental stage by the age of six. And, contrary to studies using imitation procedure (6), the spontaneous productions of children are different from those of adults: between the ages of four and six, timing of vowel and consonant in VC sequences becomes adult-like by restructuring vowel duration preceding voiceless consonants.

5. REFERENCES

PREMEANINGFUL VOCALIZATIONS OF HEARING-IMPAIRED AND NORMALLY HEARING SUBJECTS

Carol Stoel-Gammon

University of Washington
Seattle, Washington U.S.A.

ABSTRACT
The present study extends the work of Stoel-Gammon [3] by examining longitudinal samples of non-meaningful vocalizations from 10 normally hearing subjects, aged 5-18 months, and 11 hearing-impaired subjects, aged 5-39 months. Consonantal phones in the samples were phonetically transcribed and analyzed in terms of proportional occurrence of place and manner classes. Developmental trends within each group were also examined. The results show clear group differences in both place and manner of articulation. The hearing-impaired subjects evidenced a higher proportion of labial and supraglottal stops, and a lower proportion of alveolar and supraglottal stops. There were also clear trends between 8 and 22 months of age.

1. INTRODUCTION
Recent research has identified several differences between the prelinguistic development of normally hearing (NH) and hearing-impaired (HI) infants. In particular, it has been shown that the onset of canonical babbling, which typically occurs before 9 months in the hearing infant, does not occur until 12 months of age in HI subjects [2] and that the phonetic inventories of NH and HI subjects differ in size (HI inventories were smaller) and composition [3,4]. Stoel-Gammon's detailed comparison of the consonantal inventories of 11 NH and 14 HI subjects showed group differences in both place and manner of articulation of consonantal phones. Specifically, the inventories of the HI subjects contained more continuant phones and more types of labial than alveolar consonants, by comparison, the NH subjects tended to have more balanced repertoires with nearly equal numbers of labial and alveolar phones. In addition, the inventories of the HI subjects contained a higher proportion of syllabic consonants than a lower proportion of stops than the NH group. Since the study focused exclusively on consonantal inventories (i.e., on consonantal types), it provides a partial picture of the phonetic characteristics of prelinguistic vocalizations of the two groups.

The present study extends the work Stoel-Gammon [3] by examining the frequency of occurrence of each consonantal phone (i.e., analysis of consonantal tokens) and determining the proportional use of particular place and manner classes.

2. METHODS
The subjects and database for the present study are a subset of those used in the previous study by Stoel-Gammon [3]. Methodological procedures are briefly described in the following sections; for more complete descriptions, particularly of the HI subjects, readers are referred to the previous publication.

2.1. Subjects
The NH group consists of 10 subjects whose prelinguistic development was followed from around 5 months to the onset of meaningful speech, usually around 15-18 months. (These subjects are identified as N1-10 in the previous publication.) None of the HI subjects suffered from recurrent otitis media during the study.

The HI group consists of 11 subjects, aged 5-39 months, with moderate-severe sensorineural hearing loss. (These subjects are identified as YH 1.2,5,6,7 and OH 1.2,4,5,6,7 in the previous study [3]. Details regarding hearing sensitivity, age at loss, age at identification of loss and amplification are provided in that reference.) The HI subjects varied in age at onset and age at identification of hearing loss; for five subjects, data are available in the 5-18 month age range corresponding to the period of data collection for the NH subjects. The remaining six subjects were 19 months or older at the time of data collection.

2.2. Data collection
Half-hour audio recordings were collected in a sound-treated room during which parents and experimenters used eye contact and vocalizations to stimulate vocal output. To be included for analysis, a sample had to contain at least 10 speechlike utterances with a minimum of 20 consonant tokens. The maximum number of speechlike vocalizations for any one sample was set at 60.

Samples were collected from the NH subjects at approximately 6-10 week intervals. The database for this group contains a total of 44 samples with the number of samples per subject ranging from 5-6. The database for the HI group consists of 28 samples. Longitudinal data are available for eight subjects; data for the remaining three consist of a single recorded sample. 12 of the HI samples are from subjects under 18.4 months and thus overlap with the age range of the hearing group.

2.3. Data Analysis
Speechlike vocalizations of each sample were transcribed by a team of trained transcribers who worked independently and then compared analyses. Transcriptions were not changed unless a transcriber felt he or she was mistaken after relistening to the sample. Comparison of 10% of the transcriptions showed that intertranscriber agreement for place, manner and voicing of consonants exceeded 90%. For the present study, the two transcriptions of each sample were analyzed independently to determine the number of occurrences of each consonantal phone and the proportional occurrence of consonants according to traditional place and manner classes. The analysis of place of articulation was based on four categories: (1) labial, including labiodental; (2) alveolar, including interdental and palatal; (3) velar, including uvular and pharyngeal; and (4) glottal. For manner of articulation, consonants were categorized as one of the following: (1) stop; (2) fricative; (3) affricate; (4) nasal; (5) glide; (6) liquid; and (7) flap or trill. The proportion of syllabic consonants, a category which overlapped with some of the manner categories identified above, was also determined. The percentages for each place and manner category obtained from analysis of the independent transcriptions were averaged to yield a single percentage for each place and manner class for each sample.

3. RESULTS AND DISCUSSION
To provide a general picture of the phonetic characteristics of the vocalizations of subjects in each group, the overall performances of NH and HI subjects were compared. The samples were then grouped by age in order to examine developmental trends within each subject population.

3.1. General comparison
Previous studies [2,4] suggested that the vocalizations of HI and NH children showed evidence of higher proportion of glottal consonants than those of NH subjects and this was supported by the findings of the present study. Across all samples, the mean proportion of glottals for the NH group was 24.1% (SD14.8) compared with 36.6% (SD 28.3) for the HI group. As shown by the large standard deviations, there was a good deal of variance across samples; in fact, although the mean proportion for the HI samples was just over 36%, one sample contained no supraglottal tokens. Although the proportional use of glottals was higher for the HI subjects, differences in place and manner of articulation of supraglottal consonants were of even greater magnitude. Table 1 presents a set of key differences between the two groups in the use of supraglottal consonants. (Percentages in this table are based on an analysis of supraglottal consonants only, and thus represent a subset of the data.) In terms of place of articulation, the suggestion by Stoel-Gammon [3] that HI
subjects produce relatively more labial consonants and fewer alveolar consonants is borne out by the frequency of occurrence data. In the HI samples, labial consonants accounted for a much higher proportion of the data, nearly 72% of the supra-glottal consonants produced; in the NH samples, the mean proportion of labials was 42%. The figures for alveolars show the opposite trend with the proportional use by NH subjects nearly three times as high as for HI subjects (34.4% vs 12.1%). Here again, the standard deviations are quite high; part of the variance can be explained by developmental changes which are discussed below.

TABLE 1. Group comparisons: Mean occurrence of place and manner features as a proportion of supra-glottal consonants.

<table>
<thead>
<tr>
<th></th>
<th>NH</th>
<th>HI</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Lab</td>
<td>42.0</td>
<td>71.7</td>
</tr>
<tr>
<td>%Alveolar</td>
<td>34.2</td>
<td>12.1</td>
</tr>
<tr>
<td>%Stop</td>
<td>34.4</td>
<td>14.4</td>
</tr>
<tr>
<td>%Nasal</td>
<td>24.9</td>
<td>50.5</td>
</tr>
<tr>
<td>%Syllabic</td>
<td>22.8</td>
<td>43.2</td>
</tr>
</tbody>
</table>

The comparison of manner features highlights three areas in which the group samples differed markedly: the HI sample contained a much higher proportion of nasal consonants and a much lower proportion of supra-glottal stops. In addition, the HI subjects produced proportionately more syllabic consonants, many of which were nasals.

3.2 Developmental comparisons

The second type of group comparison focuses on changes in the proportional use of particular place and manner features as a function of age. NH samples were classified by age as Early (5.0-7.3 months), Mid (8.0-13.6 months) or Late (14.4-18.4 months). Table 2 presents a comparison of NH samples grouped by these age periods; only those place and manner categories which showed a change with age are shown in the table. As in the previous table, the percentages represent the relative occurrence of features of supra-glottal consonants only.

TABLE 2. NH Subjects: Place and manner of supra-glottal consonants by age.

<table>
<thead>
<tr>
<th></th>
<th>Early</th>
<th>Mid</th>
<th>Late</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Lab</td>
<td>(27.8)</td>
<td>36.7</td>
<td>32.3</td>
</tr>
<tr>
<td>%Alveolar</td>
<td>(13.7)</td>
<td>13.9</td>
<td>46.2</td>
</tr>
<tr>
<td>%Stop</td>
<td>(15.8)</td>
<td>18.0</td>
<td>43.0</td>
</tr>
<tr>
<td>%Nasal</td>
<td>(19.3)</td>
<td>14.4</td>
<td>16.3</td>
</tr>
<tr>
<td>%Syllabic</td>
<td>(21.2)</td>
<td>24.9</td>
<td>50.5</td>
</tr>
</tbody>
</table>

It can be seen that each of the features in question shows a linear increase or decrease as a function of age and that, the amount of variance for each feature tended to be highest in the Mid age range. For place of articulation, there is a marked decrease in the proportion of labial consonants and an increase in the proportion of alveolar consonants with age. In both cases, the degree of change between the Early and Mid age range remains relatively similar to the latter period indicating more uniform performance.

For manner of articulation, the mean proportional occurrence of supralabial consonants generally remains more than doubles between the Early and Mid age periods, rising from 18% to 40%, and then increasing slightly in the subsequent period to 43%. Here again, the amount of variance declines in the third period. The proportion of syllabic consonants decreases substantially with age, from nearly 50% of all supra-glottal consonants in the Early period to about 6% in the Late period.

Table 3 presents a comparison, based on analysis of supra-glottal consonants, of HI samples grouped by three age periods: Early (5.0-12.0 months), Mid (15.0-21.2 months) and Late (22.7-39.4 months). The evidence from the table shows that the developmental patterns of the HI subjects do not follow the linear trends noted for the NH group; rather, they are better described as U-shaped patterns wherein the samples in the Mid age show a marked increase or decrease in the occurrence of a sound class and the samples in the Late age period show a reversal in the direction of change.

TABLE 3. HI Subjects: Place and manner of supra-glottal consonants by age.

<table>
<thead>
<tr>
<th></th>
<th>Early</th>
<th>Mid</th>
<th>Late</th>
</tr>
</thead>
<tbody>
<tr>
<td>%Lab</td>
<td>(22.7)</td>
<td>37.2</td>
<td>50.0</td>
</tr>
<tr>
<td>%Alveolar</td>
<td>(24.1)</td>
<td>23.8</td>
<td>(SD)</td>
</tr>
<tr>
<td>%Stop</td>
<td>(11.7)</td>
<td>20.9</td>
<td>(SD)</td>
</tr>
<tr>
<td>%Nasal</td>
<td>(33.7)</td>
<td>50.0</td>
<td>(SD)</td>
</tr>
<tr>
<td>%Syllabic</td>
<td>(27.5)</td>
<td>50.0</td>
<td>(SD)</td>
</tr>
</tbody>
</table>

The mean proportion of labial consonants, for example, increased sharply between the Early to the Mid age, from a mean of 37.2% to 90.8%; in the Late age period, the mean dropped to 74.5%. A similar pattern is seen in the occurrence of alveolar which decreased from a mean of 23.8% in the Early period to 6.9% in the Mid period and then increased to 12.4% in the Late period. The proportional occurrence of supralabial stops and syllabic consonants also showed reversals in their developmental patterns.

Comparison of Tables 2 and 3 reveals that the performance of the two subject groups were remarkably similar in the samples from the youngest subjects and began increasing dissimilar with age, up to 22 months. It is not possible to make direct comparisons of HI and NH subjects over 22 months of age since the non-meaningful vocalizations of the NH subjects at this age were not analyzed. It is clear, however, that the U-shaped developmental curves in this HI samples make the production of supra-glottal consonants more similar to the NH patterns.

In sum, two major differences between the groups emerge from the analyses. First, the HI subjects produce a higher proportion of labial phones. This difference is most likely due to the fact that labials have a highly salient visual component and their articulation can be seen and imitated by babies who have little or no auditory input; alveolar consonants, by comparison, lack the visual component. Second, the HI subjects produce more nasals and syllabic consonants. It was hypothesized earlier [3] that this preference is due to the fact that these consonants provide more tactile and kineesthetic feedback than do stops which are characterized by rapid movements and short durations.

More research is needed, particularly with HI subjects at younger ages, before the hypotheses proposed here can be confirmed. By documenting phonetic patterns in one set of HI subjects, the present study provides a starting point for such research.

ACKNOWLEDGEMENT This work was supported by the National Institute of Health: grants 101-H12695 and P01-NS26521.

REFERENCES
A LONGITUDINAL STUDY OF THE SPEECH ACQUISITION OF THREE SIBLINGS DIAGNOSED AS VERBALLY DYSPRAXIC

M Tate

Amplivox Children's Speech and Hearing Centre, Cheshire, England

ABSTRACT

Developmental Verbal Dyspraxia (DVD) is a term used to denote a disorder of planning oral movements, which is present developmentally. This paper introduces an in-depth longitudinal study of three siblings diagnosed as verbally dyspraxic. The study seeks to establish characteristics of the condition and to highlight differences and similarities between the children. The study supports the notion of DVD as a syndrome in which the central phonological problem is interlinked with other language deficits and a more generalised dyspraxia.

1. INTRODUCTION

Developmental Verbal Dyspraxia (DVD) is a term which occurs in the literature and is used in clinical diagnosis, in its own right. The condition is one in which children have moderate to severe articulation defects without any apparent organic cause. However, it is not clear whether DVD is a pure disorder of the sound system, or whether it is a broader syndrome. This author had the opportunity to make a longitudinal study of three children diagnosed as having severe verbal dyspraxic problems.

2. PROCEDURE

The study [4] was retrospective and made use of tape recordings and written notes collected over a ten year period, which covered stages from early babyhood until each child had reached a high degree of spoken competence. Recordings, which have been checked for accuracy of transcription, were made at intervals of approximately 4 to 6 months, and relate mainly to conversation with adults, particularly with the caregiver and with each child's speech therapist.

3. CHARACTERISTICS OF DVD

3.1 The Existence of a Syndrome

A clear definition of DVD is hard to find. The central problem is seen to be a disorder at the speech sound level and there appear to be some essential speech symptoms [2]. It has also been suggested that children with DVD demonstrate symptoms of a still wider disorder [3]. The main speech symptoms, as described in the literature, may be summarised as:

a) inconsistency in articulated production
b) difficulty in selection and sequencing of phonological and articulatory movements
c) increasing difficulty with increasing complexity of sequences
d) altered prosodic features, and
e) difference between voluntary and involuntary movements.

There may also be an accompanying:
f) expressive language disorder
g) learning disability, and
h) general motor problems.

These features are considered with relation to the children studied.

Each child displayed signs and symptoms characteristic of DVD, with phonological difficulties as the primary feature, see Table 1.

Table 1: Details of Children Studied

<table>
<thead>
<tr>
<th>Sibling</th>
<th>Date of Birth:</th>
<th>I.Q.</th>
<th>Age of Diagnosis:</th>
<th>Severity of DVD:</th>
<th>Major difficulties:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.3.77</td>
<td>119</td>
<td>3 yrs</td>
<td>Severe</td>
<td>Phonology, Syntax</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Articulation, Clumsiness</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Auditory Memory</td>
</tr>
<tr>
<td>2</td>
<td>14.1.81</td>
<td>131</td>
<td>1 yr 10 m</td>
<td>Severe</td>
<td>Phonology, Syntax</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Articulation, Clumsiness</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Writing, Spelling</td>
</tr>
<tr>
<td>3</td>
<td>22.3.83</td>
<td>113</td>
<td>3 yrs 5 m</td>
<td>Mild</td>
<td>Phonology</td>
</tr>
</tbody>
</table>

3.2 Inconsistency in Articulated Production

The children's articulated productions could be characterised as very variable, sometimes following an adult pattern, at other times varying even within a single lexical item. Their earliest productions showed the greatest variation, and over a period of time, favoured versions could be identified. Variability was a feature of both vowel and consonant usage.

3.3 Difficulty in Selection and Sequencing

Vowels are rarely error in most children, but were very noticeable in these children's speech, although the difficulty did not lie in an inability to produce the required vowels, and early words include examples of both correct and incorrect production.

Most of the children's early words were monosyllables and many of these were open vowels. Normally developing infants use open vowels in less than 5% of their words [1], whereas in these children they accounted for up to one third of their early words.

Errors in consonant selection and some infrequent sequencing errors, accompanied vowel errors. Several normal phonological processes were identified in the children's speech, notably syllable deletion, final consonant deletion and cluster reduction, which they used, extensively until later than normal, possibly due to their articulatory difficulties. There is also
evidence of the use of some idiosyncratic processes, these are error patterns, not documented, or infrequent, in normal children, and of some chronological mismatch, where processes used in normal development co-occur with some correct production of sounds usually acquired late.

3.4 Increasing Difficulty with Increasing Complexity
Polysyllabic words created particular problems, with great difficulty occurring in the production of words of more than two syllables. Sometimes such words were shortened, in almost all cases sounds were rearranged and substitutions made. They were unable to repeat polysyllabic words even when broken down into their constituent parts. These difficulties were slow to resolve and a continuing difficulty with polysyllabic words was still evident at the end of the study.

3.5 Altered Prosodic Features
The three children's early vocalisations varied from the norm. Their vocalisations were not wide ranging, although their use of reflexive vocalisations, crying and laughing, were normal. They were quiet babies who failed to babble freely, and whose productions were limited in both character and length. A pattern of reduplicated CV syllables in babbling was present but far from striking. Most of their utterances were single syllables and lacked flow. Their early vocalisations appeared not to be progressively shaped by the auditory pattern of the adult speech around them and screaming and crying increasingly became part of their utterances. The quality of their production continued to be somewhat unpredictable, Rhythm was restricted by the use of mono-syllables and temporal delay. They used flat intonation which did not improve with increases in their phonetic inventories and the length of their utterances. The use of a deep voice, the introduction of intrusive sounds, and a preference for sounds produced at the back of the mouth, made their speech appear tense and effortful. The children all appeared to need to apply great thought and planning to their utterances.

3.5 Differences between Voluntary and Involuntary Movements
It is not clear that basic involuntary movements were entirely without difficulty, but these were much easier for them than similar actions performed on imitation or as part of speech. Tongue control exercises, for example, were more difficult, and imitation of tongue movements was only possible voluntarily after several months of speech therapy.

3.6 Expressive Language Disorder
All three children's early expressive language lagged significantly behind their comprehension. Even when their language reached an age-appropriate level, it contained widespread errors, both normal and deviant in nature. It showed a mismatch of development, containing features from a variety of stages, and also demonstrated considerable limitation in vocabulary. Particular difficulty was found in the use:
   a) Pronouns
   b) Verb tenses
   c) Prepositions
   d) Question forms and negative structures

4. Discussion
The study of these three children indicates that there exist related areas of difficulty which go beyond those which could be caused by a pure motor programming deficit. Whilst it is not possible to state that DVD cannot exist as a pure disorder of the sound system, in these children it was not confined in this way. The evidence tends to support the argument that DVD is not a pure phonological disorder, but rather that DVD is a syndrome complex, in which a severe and persistent phonological disorder is linked with other characteristics. The characteristics evidenced vary across the children, both in type and degree, but when grouped together they support a cluster of symptoms which appear also in the literature and which seem collectively to create a distinct syndrome of DVD.

Many children with DVD are not diagnosed until their speech patterns are relatively fixed, making remediation more difficult. Based on this study, it is possible that early predictors of difficulties in speech development can be found. Characteristics that may indicate that a young child's speech should be monitored include restricted early babbling, limited vocal response to stimulation, vowel errors and the common use of open vowels, variability of production and vocabulary limitations. The children's general development lends support to the existence of DVD not as an isolated and exclusive condition, but rather as one type of developmental dyspraxia in which phonological difficulties are the primary feature, interlinked with the existence of other language deficits, particularly of syntax and spelling, and accompanied by mild clumsiness and poor fine co-ordination in other areas, although these may be varied in type and degree.

5. References
EXAMINATION OF LANGUAGE-SPECIFIC INFLUENCES IN INFANTS' DISCRIMINATION OF PROSODIC CATEGORIES

C. T. Best¹, A. G. LeVitt² & G. W. McRoberts³

Haskins Laboratories, New Haven CT 06511, USA and ¹Wesleyan Univ., Middletown CT; ²Wellesley College, Wellesley MA; ³Stanford Univ., Palo Alto CA

ABSTRACT
Language-specific effects in perception of segmental contrasts appear by 10-12 months. Recent studies with connected speech suggest earlier emergence of sensitivity to some language-specific prosodic properties, but they have not examined linguistic prosodic contrasts. We tested 6-8 and 10-12 month olds on a discourse prosody contrast (question-statement) in native and non-native languages. Category discrimination was significant for native, nearly so for non-native speech. Separate analyses found younger infants discriminated in both languages, older infants in neither, failing to support language-specific perception of this prosodic contrast.

1. INTRODUCTION
To acquire language, the infant must learn to recognize that certain sound patterns recur in native speech, whereas others do not. Adults show language-specific attunement in perception of phoneme contrasts, often finding it initially difficult to discriminate non-native segmental distinctions [10, 11, 15]. But infants under 8 months discriminate both native and non-native contrasts. Difficulty distinguishing non-native contrasts appears by 10-12 months [2, 3, 14].

Infants must also learn the prosodic characteristics of the native language. Indeed, it has been argued that infants become attuned earlier to prosodic than segmental properties [7, 9]. Numerous recent findings appear consistent with this claim. Infants from 5 months to as young as 1-2 days prefer infant directed speech (IDS) over adult-directed speech [6], and can discriminate native from non-native connected speech [1, 12], even when segmental content is removed from the F0 contours. Other language-specific effects on prosodic perception appear by 6-11 months [5, 6, 8]. Even in utero exposure to mother's voice can affect newborn preferences for familiar patterns in her speech [4, 5]. Thus, many experience-based effects on prosodic perception are found earlier than the 10-12 month reorganization for segmental contrasts. Yet direct comparison of the prosodic and segmental findings is problematic. Whereas the segmental studies tested discrimination of phonemic contrasts, the prosodic studies have examined responses to broad prosodic patterns and have not tested linguistic contrasts. Therefore, we examined infants' discrimination of a prosodic contrast in native vs. non-native speech.

The question-statement contrast is a discourse distinction whose prosodic patterns may be within the infant's reach. Discourse prosody may help infants discover certain pragmatic distinctions without lexical knowledge. That is, interrogative intonation indicates the response is expected from the listener, while declarative intonation indicates a comment directed toward the listener.

Although questions are often marked by final F0 rise, and statements by final fall, these characteristics are not entirely consistent, particularly in IDS [7]. For example, Spanish questions show fairly consistent final rise, but English wh-questions show an earlier pitch peak and final F0 decline, while Spanish and French continuation statements show final rise. Thus, recognizing that diverse utterances converge or contrast on discourse categories requires detecting abstract, language-specific commonalities among varying F0 patterns. For this reason, we tested infants' recognition of native vs. non-native prosodic contrasts across multiple questions and statements.

2. METHOD
2.1 Subjects
Monolingual English-learning American 6-8 and 10-12 month olds were tested on prosodic contrasts in English and Spanish. At each age, eight infants completed a categorical-change condition, eight an arbitrary-change condition.

2.2 Stimulus Materials
Three questions and three statements (exclamatory in IDS), all seven syllables long, were matched for content in English and Spanish: What a beautiful baby! (Qué bebe tan guapo!); You are such a great, big boy! (Beres un niño grande!); My beautiful little doll! (Mi muñequita linda!). Who is this little fellow? (Quién es este niño?); How are you doing today? (Y como estás hoy?); And whose sweet baby are you? (De quién es este bebé?). A female speaker of American English, and one of Mexican Spanish, produced multiple IDS phrases as though to a young infant.

One token per sentence was selected to provide comparable between-sentence duration, loudness, F0 level and range. Within-language differences in duration and loudness were reduced by waveform editing. Figures 1 and 2 show the F0 contours for the final set in each language. F0 range was larger for questions than statements; the difference was more extreme for English. Only the Spanish questions showed final rise.

2.3 Procedure
Discrimination was tested in a habituation procedure that employed a conditioned visual fixation response [3]. Subjects in each condition received two tests, one per language. In the categorical condition, infants were initially presented with randomly-ordered repetitions of either the question or the statement in a given language, contingent on their fixation of a target slide. Once fixations fell below the habituation criterion (two consecutive trials at less than 50% of the mean for the first two trials), audio presentations were shifted to the opposing discourse category in the same language. Infants in the arbitrary condition received a change from one within-language mixture of questions and statements to another. The categorical shift should be discriminated better than the arbitrary shift if infants show perceptual constancy for prosodic properties shared by the diverse items within
each discourse category. A language-specific influence would be evident if categorical discrimination were better for native than for non-native sentences.

3. RESULTS

Mean fixation times in the last two trials before the stimulus shift were compared to mean fixation times in the first two trials following the shift, in an Age x Language x Condition (categorical vs. arbitrary) x Shift (pre vs. post) ANOVA. Fixations were longer post-shift than pre-shift \( F(1,28) = 15.04, p < .006 \), indicating overall discrimination. Simple effect tests found discrimination only in the categorical condition \( F(1,39) = 10.17, p < .001 \), which was significant for English \( F(1,14) = 10.96, p < .005 \) and nearly so for Spanish \( p = .058 \). The Language x Condition effect \( F(1,28) = 4.66, p < .04 \) found that fixation times were highest in the English categorical condition, lowest in the English arbitrary condition. A nearly-significant Age x Condition x Language interaction \( p = .057 \) suggested differences in younger and older infants’ response pattern.

We therefore tested the possibility that language-specific effects were reliable for only one age group, as in previous findings that language-specific effects in perception of segmental contrasts appear around 10–12 months. However, separate analyses failed to support language-specific effects for the prosodic contrast at either age. The 6–8 month olds discriminated the category change, but the arbitrary change, in both English \( F(1,17) = 8.29, p < .024 \) and Spanish \( F(1,17) = 14.42, p < .007 \).

The 10–12 month olds failed with both individual languages, showing marginal categorical discrimination overall \( p > .08 \). Figure 3 shows these post-shift recovery patterns.

4. DISCUSSION

The present task required that the infants detect abstract commonalities among the diverse sentences within each category. The overall ANOVA suggested that, across ages, infants distinguished the discourse categories of question vs. statement, but not between arbitrary groupings of the same sentences. Further research will be needed to determine the prosodic properties that guide infants’ perception of these categories. The Spanish questions were quite similar in their F0 contours, all showing final rise, which differed from the consistent F0 decline of the statements. But the F0 contours in each English category were quite variable, and were not distinguished by final rise vs. fall. Nonetheless, across ages the infants discriminated the English with better reliability than the Spanish categorical change, suggesting that final rise/fall was not the critical perceptual feature for them. Both languages had greater F0 range in questions than in statements; this property may have been more salient to the infants, either in both languages or at least in English.

The overall discrimination was marginal across languages. The IDS properties of the sentences themselves suggest a possible clue to the older infants’ difficulty: they were addressed to much younger infants. Speech to infants near the end of the first year often contains prosodic features that emphasize referential objects to people, whereas that to very young infants comments primarily on the infant’s state or activities (without emphatic references to objects [13]). Perhaps 10–12 month olds would discriminate this prosodic contrast if it were carried in age-appropriate utterances. Alternatively, older infants may be more sensitive to prosodic properties, and more focused on segmental and/or lexical information, than are younger infants.

This study provided little evidence for further attunement to native prosodic contrasts than to segmental contrasts. On the contrary, the 10–12 month reorganization in perception of non-native segmental discourse appeared to be preceded or even paralleled by analog reorganization in the perception of this linguistic prosodic contrast.

5. ACKNOWLEDGMENT

Supported by NICHD grant HD-01994 to Haskins Laboratories and NICDD grant to Dr. Ryan, the first author.

6. REFERENCES


ARTICULATORY ORGANIZATION OF EARLY WORDS: FROM SYLLABLE TO PHONEME

Elizabeth W. Goodell and Michael Studdert-Kennedy

University of Connecticut and Haskins Laboratories, Trinity College, Hartford, CT, Yale University, New Haven, CT

ABSTRACT

Evidence that children's initial units of phonological contrasts are words or short formulaic phrases rather than phonemes or features invites the hypothesis that the initial domain of articulatory (or gestural) organization may also be larger than the phoneme. The present study investigates the development of intrasyllabic gestural overlap in fricative-vowel syllables between the ages of 22 and 32 months. Results indicate that children at both ages display more gestural overlap than adults.

1. INTRODUCTION

Studies of early phonological development have typically taken abstract linguistic units (phonemes, features) as undifferentiated, phonological primitives, and have implicitly, or explicitly, attributed a functional role to these units in the perceptual representation and articulatory organization of a child's early words [1]. Encouraging the notions that: (1) the units of linguistic contrast in a child's early speech are not phonemes and features, but words, or formulaic phrases, consisting of one or a few syllables [2]; (2) the initial units of articulatory organization are gestural routes extending over a word or phrase [3, 7]; (3) phonemes and their featural descriptors emerge from syllables by gradual differentiation of consonantal and vocalic oral gestures [3, 6]. Results consistent with this account have come from a study of fricative-vowel coarticulation (or gestural overlap) in young children and adults, in which 3-year-old children uttering fricative-vowel syllables displayed significantly more gestural overlap between fricative and vowel than older children and adults [4]. The present 10-month longitudinal study extends the preceding investigation to younger ages: 22 and 32 months.

2. METHOD

The subjects were six girls (mean age=22 months, mean MLU=1.36, at beginning of study) and six adult females. The test utterances were designed to investigate fricative-vowel coarticulation in CCVC contexts, similar to previous studies [cf. 4, 5]. The utterance types were three nonsense syllables: /uwa/, /al/, and /au/. The vowels, [a, i, u], were chosen because they occupy extreme points in the vowel space so that the vowels of fricative-vowel syllables were anticipated in the fricatives, differences in the lingual front-back dimension, as indicated by estimates of the fricative second formant (F2), should be apparent.

The children's data were collected in the first and tenth months during half hour sessions with the experimenter in the child's home. As many utterances as possible were elicited through games with stuffed animals. Out of a total of 234 children's utterances, the resulting number of acceptable utterances of each type for each child ranged from 2 to 20, with a mean of 6.5. Nine utterances from the children's data were excluded due to background noise or lack of formant structure in V1. The adults produced 6 utterances of each type in random order. No adult responses were excluded.

All tokens were digitized at a 20-kHz sampling rate on a VAX 780 computer, and a waveform editing and display system was used to measure the duration of the first fricative and vowel. Five locations for estimating formant frequencies were then chosen: (a) the midpoint of the initial fricative (1/2 fric), (b) the onset of voicing for the first vowel, (c) the midpoint between (a) and (b) (3/4 fric), (d) the midpoint of the first vowel (1/2 vowel) and (e) the midpoint between (b) and (d) (1/4 vowel). Estimation of the center frequencies of the second formants were made at these five locations from Discrete Fourier Transform spectra, computed with a 25.6 usec. Hamming window and a 3.2 msec. slide between windows. F2 estimates could not be made at both points in the fricative of every token: 54% of the adult tokens permitted F2 estimates at 1/2 fric, 7/8 at 3/4 fric; 76% of the children's data permitted estimates at 1/2 fric, 85% at 3/4 fric. Estimation of the center frequencies for the first formants were made at the last three points. All vocalic formant estimates were made by finding the highest amplitude harmonic in the region of the given fricant at a given location and computing the weighted mean of this harmonic and the harmonics immediately above and below it.

3. RESULTS

3.1 Gestural Overlap in the Fricative

Figure 1a, b, c displays the mean estimated formant paths for children, 32-month-olds, and 22-month-olds respectively. In Figure 1a (adults) the F2 measurements at 1/2 fric are practically the same as for all three vowels. At 3/4 fric a vowel /a/ distinction begins to appear with differences of about 300 Hz between F2 values before [a] and the back vowels. Finally, at 1/2 fric, a vowel /u/ space has emerged, in which [u] has much higher F2 values than [a] [cf. 5].

For the 32-month-olds (Figure 1b), substantial anticipatory gestural overlap is apparent in the formant values at 1/2 fric and 3/4 fric with differences of roughly 200 to 500 Hz between the values preceding the different vowels. The front and back vowel formant paths continue to diverge, but at 1/2 vowel the F2 estimates are only slightly higher for [u] than for [a], indicating that the children are relying largely on tongue height to distinguish the vowels (see F1 values).

Finally, the 22-month-olds (Figure 1c) display much the same degree of gestural overlap as their older selves in the front-back dimension, as evidenced by the different formant values for [a, u] vs [I] at both 1/2 fric and 3/4 fric. However, unlike their older selves, they do not differentiate the fricatives more for [u] and [a], and the final values of F2 at 1/2 vowel for [u] and [a] reverse the pattern observed in the adults. Both the
latter effects arise from an overall higher formant peak for [sa] at the younger age. (See below under Gestural Overlap in the Vowel.)

As an index of gestural overlap permitting comparison across groups, self-normalization ratios were formed: the fricative F2 values for [i] were placed over the fricative F2 values for [u] and [a] at the 1/2 fric and 3/4 fric measurement points. This ratio is an index of the degree of gestural anticipation; if the value is 1, there is no difference between fricative formant measurements before the two vowels, indicating no anticipation of the following vowel. The farther the value from 1, the greater the anticipation of the vowel.

Table 1 lists the mean ratios for adults and children at 1/2 fric and 3/4 fric points. At 1/2 fric F2 values are significantly different before [u] than before both [a] and [u] for the 22-month-olds, before [i] than before [a] but, due to a single deviant, not before [u] for the 32-month-olds. There are no effects of vowel for the adults at this point. At 3/4 fric F2 values are significantly different before [i] than before both [a] and [u] for all except the 32-month-olds before [u] (again due to a single deviant subject).

Table 1. Amount of gestural anticipation at two points in the fricative, indexed by mean ratios of fricative F2 values before [i] to fricative F2 values before [u] and [a]. An index significantly greater than 1.00 indicates a significant degree of gestural anticipation. * p < .005, one-tailed test.

<table>
<thead>
<tr>
<th>Measurement Point</th>
<th>1/2 Fricative</th>
<th>3/4 Fricative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adults</td>
<td>1.00</td>
<td>1.04</td>
</tr>
<tr>
<td>32-month-olds</td>
<td>1.15</td>
<td>1.12</td>
</tr>
<tr>
<td>22-month-olds</td>
<td>1.15</td>
<td>1.13</td>
</tr>
</tbody>
</table>

3.2 Gestural Overlap in the Vowel

As noted above, the relative positions of [u] and [a] at 1/2 vowel differ for the three groups. These differences are displayed in Figure 2. Notice that in the adults, F2 is higher for [u] than for [a] by about 250 Hz, in accord with [5], perhaps indicating a more forward consonant location for [u] than for [a]. For the 32-month-olds, F2 for [u] and [a] are higher for [u] than for [a] by about 70 Hz, while for the 22-month-olds F2 is higher for [a] than for [u] by about 350 Hz, the reverse of the adults. Lines connecting tokens in Figure 2 illustrate the a-g group differences.

We may now ask concerning the adults: Is the relatively higher F2 for [u] due to overlap of the vocalic gesture with the gestures of the surrounding alveolar sibils? To answer this question, more data were collected from the adult subjects. In addition to repeating the original fricative stimulus items [sa], [sas], and [asu], 6 times each, subjects also produced 6 repetitions each of [zh], [tad], [tdad], [hihi], [haah], and [hulu]. The means for the first vowels in these contexts are also given in Figure 2. Orthogonal comparisons reveal that [zh] and [tad] do not significantly differ from each other, [F(1,409), p < .2415], but do differ significantly from [hihi], [F = 39.712, p < .0001]. Apparently then [u] is articulated further forward if bracketed by alveolar stops or fricatives than if bracketed by the articulatorily neutral [h], while the position for [a] stays the same in both contexts. This result is consistent with the proposal in [5] that overlap of C and V gestures in adults is facilitated if C and V tongue heights are compatible, but are blocked if they are not.

We were not able to collect more data for acoustic analysis from the children. However, the children's original utterances were transcribed independently by two colleagues. It was discovered that many of the 22-month-olds' tokens for [a] were somewhat fronted and raised, e.g. [seyso] instead of [yas]. Evidently the 22-month-olds were not able to block gestural overlap of the low vowel [a] preceding and following [s], as were the adults and, to a fair extent, the 32-month-olds.

3.3 Durations

Children's utterances are often longer than adults'. The mean durations for fricative 1, vowel 1, and syllable 1 were therefore compared, by analysis of variance. There were no significant interactions with, or effects of, age. Accordingly, none of the differences among groups reported above can be attributed to differences in rate of speaking.

4. DISCUSSION

Consistent with the results of [4] for older children, the present study found a significant tendency for 22- and 32-month-old children to anticipate the fronting location of the vowel earlier in the fricative of a fricative-vowel syllable than adults. Two observations suggest that this result does not reflect "planned" coarticulation: (1) the greater difference at 22 months between fricative F2s for [u] and [a] at 1/2 fric than between F2s for [u] and [a] themselves at 1/2 vowel; (2) the tendency at 22 months to front and raise the low back vowel [a] in the context of preceding and following [s]. These results suggest not "planned" coarticulation, but an inability easily to differentiate and control a rapid sequence of diverse tongue gestures. This interpretation is consistent with the hypothesis that consonants and vowels emerge as stable units of articulatory control in children by differentiation of the closing and opening gestures of the canonical syllable [cf. 7]. Such an account obviates the necessity for positing phonotopes, or their featural descriptors, as underrived, phonological primitives.

Acknowledgement: Preparation of this paper was supported in part by NICHD Grant HD-01994 and DC-05403 to Haskins Laboratories, 270 Crown St., New Haven, CT, USA.

5. REFERENCES

RHYTHMIC PHENOMENA
IN A CHILD'S BABBLING AND ONE-WORD SENTENCES

M. Kohno and T. Taushima

Kobe City University of Foreign Studies
Kobe, Japan.

ABSTRACT
A baby's babbling and one-word sentences, in total 2534 utterances, were tape-recorded over a four-week period when she was at the age of 1:6-1:7, and 513 examples randomly selected from the recorded data were acoustically analyzed. It was found that 1) babbling plays a ground-breaking role for producing one-word utterances—they reveal very similar phonetic phenomena, and 2) consonant articulation is one of the most important factors to control rhythm. In addition, the following facts were also established: 1) repetition-of-two-syllable-type babbling (e.g. bakobako--) is uttered in the midst of long-short timing alternation, while simple one-syllable-repetition-type babbling (e.g. tata-ta--) reveals no distinguishable rhythmic pattern, 2) acquisition of isochronism of morae is far faster than that of syllables. 3) interestress intervals between syllables in both babbling and one-word utterances became greatly lengthened just before the period in which vocabulary abruptly increases.

1. DATA COLLECTION AND DATA ANALYSIS
The subject is a one-and-half year old Japanese female child, who has no known abnormalities. She was born and has been brought up in a Tokyo dialect area. Her utterances were recorded for about one month from March 9 to April 8, 1968, which corresponded to the period of 1:6 to 1:7 years of age. This period coincided with her single word utterance stage. The recording was done by the use of wireless microphone. Panasonic RD-53 stitched into the neck of her clothes, which was electrically connected with a cassette tape recorder, Victor VD System RC-X-5 or Aiwa SW 77. The subject's vocabulary abruptly increased at about 78 weeks of her age (March 9) from about 70 words to 200 words and therefore the whole period was divided into two periods before and after March 9 as 'early' vs. 'late' periods, respectively. This is the work that Ingram and Menou et al. [1][3] took. Each period was again divided into two sections before and after March 25 and April 6, because of the simple reason that the recording happened to be suspended for several days before these dates. All the recorded materials, therefore, were chronologically divided into two periods and four sections.

The recorded materials were then acoustically analyzed by interactive Laboratory System (I LS) run by Micro POP 11/73, AD Conversion Box, but Yokokawa Electo- Oscillograph, type 3001, connected with Amplifier 3125, was also used supplementarily.

2. ABOUT BABBINGS
2.1. Syllabic Constitution
    Intervals among voice-onset points of syllables (inter-stress intervals, ISI hereinafter) especially of syllables which have plosive-like sounds as consonant partners of CV constructions, were instrumentally measured. The numbers of utterances thus measured were 130 groups, 245 successes and 864 syllables. This means that the authors analyzed the most typical syllables of babblings. We can classify syllable structure into five groups according to the syllable repetition-alternate repetition of two different syllables (authentic and para types), simple repetition of mono-syllable (authentic and para types) and non-repetition type. Table 1 shows distributions of occurrences of these types classified by the syllable number of babbling succession. We can see here that the two-syllable repetition types are produced far more than the mono-syllable repetition type in this stage of language acquisition, but according to Stark [6] and Oller [4], the latter types are more popular than the former ones in the pre-single word stage. The mono-syllable simple repetition type of babbling (Type c) occurred 29 in total in our data (Table 1), but 23 of them appeared in the early period and only 6 in the late period. As for the two-syllable alternate repetition type (Type a), on the other hand, 55 out of 72 utterances occurred in the late period and 17 in the early one. These facts support the above observations of Stark and Oller [4][5] and lead us to the fact that Type a is more typical in the one-word utterance stage. All the non-repetition type babblings took place in the early period without exception—random, nonsystematic utterance also constitutes a characteristic feature of the early period.

2.2. Timing Control System in Babbling
There was found some regularity in ISIs among syllables in two syllable alternate repetition type, but no regularity at all in simple repetition of mono-syllable babbling.

Table 1 Syllable constitution of Babbling

<table>
<thead>
<tr>
<th>Type</th>
<th>7 syl.</th>
<th>6 syl.</th>
<th>5 syl.</th>
<th>4 syl.</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>a 2-syllable alternate repetition</td>
<td>41</td>
<td>4</td>
<td>23</td>
<td>14</td>
<td>[bakohako- bakohako-]</td>
</tr>
<tr>
<td>b 2-syll. alternate repetition in part</td>
<td>7</td>
<td>16</td>
<td>2</td>
<td>[bakobako- bakobako-]</td>
<td></td>
</tr>
<tr>
<td>c mono-syllable simple repetition</td>
<td>8</td>
<td>12</td>
<td>0</td>
<td>[tata-tata-]</td>
<td></td>
</tr>
<tr>
<td>d mono-syll. simple repetition in part</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>[tata-tata-]</td>
<td></td>
</tr>
<tr>
<td>e no repetition</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td>[piikoide-]</td>
</tr>
</tbody>
</table>

Table 2 ISIs among syllables in babbling

<table>
<thead>
<tr>
<th>Type</th>
<th>4 syllables</th>
<th>3 syllables</th>
<th>2 syllables</th>
<th>1 syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>a 2-syllable alternate repetition</td>
<td>41</td>
<td>4</td>
<td>23</td>
<td>14</td>
</tr>
<tr>
<td>b 2-syll. alternate repetition in part</td>
<td>7</td>
<td>16</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>c mono-syllable simple repetition</td>
<td>8</td>
<td>12</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>d mono-syll. simple repetition in part</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>e no repetition</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

In Table 2 which shows means of ISIs (ms) among syllables, S.D. and auto-correlations of the adjacent ISIs in each type of babbling, we can see that the values of auto-correlations in two syllable babblings are all negative, while those in mono-syllable babblings are positive in such cases as syllables, whose absolute value is very small and of seven syllable babblings. The negative auto-correlation, if its absolute value is large enough, may suggest that the ISIs of the syllables occur more or less in long-short alternation but the positive one shows no such a regularity. We should notice that the seven mono-syllable babblings which show rather high negative auto-correlation in Table 2, contrary to other mono-syllable babblings which occur in the late period, especially in Session II, except that one example occurred in Session III. This kind
of babbling therefore, despite its similarity in form with the babblings which appear in pre-single word utterance stage, may play the same role as two-syllable alternate repetition type of babbling.

The same was the syllable number of one word sentences appeared in all the recorded data. Interestingly, these numbers also coincide with the syllable numbers (72%) of perceptual sense unit (cf. [4]), that is, the chunking unit of utterance which is holistically perceived with its meaning and stored in memory in an unprocessed form in the process of listening comprehension [2].

Table 3 Recorded and Analyzed Data (single word utterances)

<table>
<thead>
<tr>
<th>Word</th>
<th>Early</th>
<th>Late</th>
<th>Subject's Reaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recorded Times</td>
<td>236</td>
<td>644</td>
<td>350</td>
</tr>
<tr>
<td>Analyzed Times</td>
<td>268</td>
<td>101</td>
<td>197</td>
</tr>
</tbody>
</table>

Table 4 Chronological Changes of ISIs (single word utterances)

<table>
<thead>
<tr>
<th>Session</th>
<th>2-2 mean</th>
<th>3-3 mean</th>
<th>AVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early</td>
<td>27.7</td>
<td>31.8</td>
<td>30.1</td>
</tr>
<tr>
<td>Late</td>
<td>27.7</td>
<td>31.8</td>
<td>30.1</td>
</tr>
<tr>
<td>S.D.</td>
<td>2.6</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>N</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 5 Comparison of Mora Lengths

<table>
<thead>
<tr>
<th>Type</th>
<th>Early</th>
<th>Late</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-2</td>
<td>2.2</td>
<td>2.2</td>
</tr>
<tr>
<td>3-3</td>
<td>2.3</td>
<td>2.3</td>
</tr>
<tr>
<td>2-2</td>
<td>2.2</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Fig. 2 illustrates the chronological change of the ISIs in single word utterances, and for comparison, the behavior of ISIs in babbling is also shown in the thick line. We can also see here amazing similarity between the two modes of sound production. As shown in the shortest intervals in Sessions III, IV and the longest intervals, respectively. More precise observation however, makes it clear that 2-2 words shape a pattern marked by 'dakko' (hold it in your hands), for instance, in the form 'da' to 'ko' than 'cho' to 'ko' (increase 'ko'). This suggests that the infant already notices the existence of moka in the infant's Japanese timing system, but this is not clear in Sessions III, IV and V. Has also been subject in the world, mastered the Japanese mora system? In order to make it clear, we carried out the following investigations.

As shown in Table 5, the ISIs in 2-2 words were significantly longer than the ones in 2-3 or 3-3 words (p<0.01) not only in the early period but also in the late period (Table 5). Throughout the periods from the early to the late periods, the means of ISIs in 2-2 words were 457.3ms and the ones in 2-3 words were 348.84ms and statistical significance at the level of p<0.01 was also detected.

We asked a Japanese adult, a university student, on the other hand, to say 'kabu', 'kou', 'aka', therefore, that the same timing control mechanism is working in babbling and one-word utterances.
L'INTONATION DE QUESTION DANS LE LANGAGE EMERGENT

Gabrielle KONOPCZYNSKI
Laboratoire de Phonétique, Besançon, France

La question figure parmi les fonctions du langage qui a le plus intéressé les psychologues, et spéciallement, ceux qui se sont occupés de la psychologie de l'enfant.

Le problème de la question est également un problème de linguistique, car elle est au cœur des problèmes de la syntaxe et de la sémantique, ainsi que de la pragmatique.

En effet, la question est un phénomène qui se manifeste à tous les stades du développement de l'enfant, de la naissance au moment de l'entrée à l'école.

Le but de ce texte est de présenter les résultats d'une étude qui a été menée dans le cadre du Laboratoire de Phonétique de Besançon, dans le but de comprendre mieux la façon dont les enfants posent des questions et comment ils utilisent l'intonation pour rendre leur question expressive.

1. METHODE

La méthode utilisée a été la méthode de l'observation, où l'enfant est observé dans ses interactions avec les autres, en particulier avec ses parents et ses frères et sœurs.

Les enregistrements ont été réalisés dans des situations naturelles, en particulier dans des situations d'interaction, et ont été analysés par un logiciel spécialisé en phonétique.

2. RESULTATS

2.1. Période charnière (9-11 ans)

Les résultats de cette étude montrent que l'intonation de question dans le langage enfant est un phénomène complexe, qui dépend de plusieurs facteurs, tels que l'âge, le sexe, et le contexte social.

Dans cette période, les enfants ont une maîtrise plus avancee de l'intonation de question, et l'intonation de question est utilisée de manière plus délibérée et plus explicitement pour marquer la question.

Les résultats montrent également que l'intonation de question est un phénomène qui se développe progressivement, et qui est influencé par les interactions sociales et les environnements culturels.

On peut dire que l'intonation de question est un phénomène qui se développe progressiveallment, et qui est influencé par les interactions sociales et les environnements culturels.
néanmoins avant la fin de la première année, sans que le langage articulé soit présent.

2.2. Entre 12 et 24 mois.

2.2.1. Les questions émises en Proto-Langage.

Ici, les six sujets sont pris en compte. Leurs questions sont plus marquées à la fois 9 et 10; car si mens plus haut dans la tésiture (niveau 4-5 jusqu'à 900Hz, cf. [6]), elles dépassent en l'essentiel l'appel, qui ont une courbe ascendante analogique, mais leur intensity est plus faible ; la courbe d'intensité, qui est toujours parallèle au F0, sauf une rapide chute finale, dépasse rarement 40 dB. Beaucoup de ces questions sont monosyllabiques, de type [æ?], de durée brève (M. = 255 ms, extrêmes 140-450 ms), alors que les vocoïdes à fonction non monosyllabique du Jassia sont toujours très longs (M. = 967 ms, extrêmes jusqu'à 8530 ms).

2.2.2. Les questions articulées sans mots ouil.

Une distinction s'impose à l'intérieur de cette catégorie entre questions marquées uniquement par la mélodie et celles articulées par un mot-outil. Les premières sont les seules attestées jusqu'au vers 20 mois, âge auquel commencent à apparaître les mots interrogatifs qui sont dans l'ordre : [es] et ses diverses formes, [ə] = oui, [komaj] = comment (22, ou seul exemple chez un sujet). Les questions sans mots ouif sont formées essentiellement d'énoncés [i] ou [i]tisylabiques, représentant des objets ou des actions dont l'enfant cherche à connaître le nom. La faible intensité est soit la forme simple, soit le mot précédé de [e], de [e] ou de [e] formant un ensemble dont le statut est difficile à déterminer ; encore monosyllabique ou déjà combinaison de deux éléments ! Souvent ce sont des formules figées, acquises globalement. Le questionnement est généralement accompagné, soit d'un geste de pointage vers l'objet, soit d'un regard interrogatif vers l'adulte.

Les caractéristiques de ces questions articulées sont résumées dans le tableau ci-dessous. On notera leur teseur élevée et l'étendue de leur glissando; forte en chiffres absolus, elle n'est pas aussi importante qu'on pourrait le penser ; le glissando des énoncés phonétiques est quelquefois plus prononcé. L'apparition du mot permet de réduire la redondance : le F0 baisse et l'adulte répète les mots en restreint. Très souvent, dès qu'il a obtenu une réponse de l'adulte, l'enfant oppose à la forme interrogative la forme énonciative ou implicative du même mot. On a par exemple :

**ENFANT**

19 ; c'est chien ! (6)
- chien !

22 ; Sophie debout dans sa baie-sno, regardant sa mère :
- assis ! (6)
- assis !

Dans ce cas, c'est toujours l'enfant qui initialise l'échange.

Dans deux autres situations, bien différentes de celle que nous venons d'étudier, l'enfant prononce successivement la forme interrogative, puis la forme énonciative. Par exemple, c'est l'adulte qui initialise le dialogue en disant un mot quelque, généralement désignation d'un objet (c'est un...), ou d'une action (on va jouer...), qui parait entendre ce mot pour la première fois, le répète d'abord sur un ton ascendant, comme s'il demandait confirmation, puis sur un ton descendant. Cette stratégie, très fréquente, semble être un moyen d'approvisionnement du lexique. Ces diverses formes ascendants sont beaucoup plus marquées que les questions habituelles ; c'est pourquoi il nous paraît difficile de les appeler "questions-échos" comme le propose BOYSSON-BARDIES [1] dans leur étude du babillage tare. Voici les caractéristiques fréquentielles de ces énoncés : Fo initial: 425 Hz (extrêmes : 350-500 Hz), Fo final 630 Hz (extrêmes : 500-850 Hz). Les auditeurs y voient généralement une question surprise. Les formes descendantes, en revanche, sont... à côté toute discussion, si l'adulte a répondu favorise. L. MENN signale une stratégie identique chez Jacob vers 17 mois.

Le second cas, également attesté chez tous les enfants suivis, est plus curieux. La situation est apparentement celle d'une interrelation : regarder avec l'enfant un catalogue. A 14 mois, l'adulte mère la danse ; la participation verbale de l'enfant est essentiellement de type mélodique ou onomatopéique. Vers 18-20 mois, il va devenir, mais l'enfant répète les mots qu'il se servait de la stratégie décrite ci-dessus. Enfin, il finit par jouer lui-même au jeu des questions-réponses : montrant un objet, il dit : "Oui, ma mère, j'ouvre !" et si l'adulte ne sert pas d'intervenant, mais simplement d'oreille réceptrice. Il semblait que ce soit là une fausse question, plutôt demande de confirmation, ou forme d'interrogation, tout comme l'est la descente peu marquée pour la partie énonciative. Nous avons relevé ce même comportement chez des enfants de six ans qui devaient dire le nom d'objets représentés sur des images. Souvent les moins les moins bien connus étaient prononcés légèrement ascendants ou peu descendants ou plutôt que l'items connu étaient émis nettement descendants. Il est intéressant d'interpréter ces "fausses" en termes, semi-énonciatifs ou énonciatifs, comme deux phases successives, la première phase servant de point de repère situationnel à l'autre et formant le cadre dans lequel la seconde est assurée ou éventuellement remise en question (CULIOIL).

Le comparatif avec des questions de même type dans le langage adulte montre des divergences sensibles. Si la forme des contours est semblable, chez l'adulte, l'étendue du glissando, qui traverse généralement deux niveaux, joue un rôle plus grand que le niveau dans lequel se situe l'énoncé (ROSSI & AL.7). Chez l'enfant au contraire il semblerait que le trait essentiel des interrogatifs soit un délai de la voix vers les zones aigus. L'utilisation des divers niveaux de la tésiture à des fins linguistiques apparaît clairement ici.

Toutes les questions sans mot-outil sont de type "interrogation botte" (FES-NO, qui appuient, non pas une information, mais une simple réponse par oui ou non. Il n'en va pas de même pour la catégorie introduite par un mot ouité, de type "interrogation partielle" qui attend une réponse plus complète. Il semblerait que l'enfant acquière ce second mode de questionnement seulement quand il est en mesure de comprendre une réponse plus élaborée que le simple acquiescement ou la pure négation.

Quelles que soient les nuances présentes dans les diverses formes éditées, il est clair que le questionnement avec une seule nodologie a un rendement maximal dans la période des premiers mots. Le trait commun à toutes ces questions métaphoriques, outre leur contour ascendant, est le niveau de voix dans lequel se situe la voix, avec un F-M toujours supérieur à 470 Hz, et une culmination des énonces dans le haut du niveau 4 ou dans le niveau 5. Ainsi les interrogatives ont le F0 le plus élevé de toutes les classes d'énoncés.

---

**TABLEAU COMPARATIF**

**QUESTIONS EN PROTO-LANGUE**

**ARTICLES**

| Forme du contour | ascendant | ascendant
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>M. Fo initial</td>
<td>406 Hz</td>
<td>459 Hz</td>
</tr>
<tr>
<td>Min. Fo</td>
<td>230 Hz</td>
<td>274 Hz</td>
</tr>
<tr>
<td>M. Fo final</td>
<td>499 Hz</td>
<td>576 Hz</td>
</tr>
<tr>
<td>Max. Fo final</td>
<td>720 Hz</td>
<td>835 Hz</td>
</tr>
</tbody>
</table>

( M. = Moyenne)

CONTEMPORARY CZECH PRONUNCIATION:  
A DATABASE STUDY

P. Janota and Z. Palková

Charles University, Prague, Czechoslovakia

ABSTRACT

Recordings of identical texts spoken by young Czech speakers, students of approximately the same age, were auditory analyzed by experienced listeners. A database structure for storing results of the auditory analysis was handled by appropriate search programs and the results of the searches were then computed transferred into tables and graphs and interpreted. Main results concerning the contemporary Czech pronunciation are presented and discussed.

One of the main tasks of phonetic departments is to describe and to analyse the current state of the vernacular language on the sound level. We have chosen the following methodic approach to evaluate the actual, existing pronunciation of the Czech language:

a) - the speech material to be analysed consisted of two short passages to be read, one easy and the other one difficult both lexically and syntactically, and a section of free narrative speech; the reading material consisted of 1) a short piece of text specially prepared for this purpose and 2) an authentic passage of prose text. The total contents of the text was 462 speech sounds (182 vowels and 280 consonants). Two minutes of free speech, recorded at the same session, were not used for the present database.

b) - several groups of rather explicitly defined speakers were recorded on tape: the first year students of Czech at the Philosophical Faculty of Charles University in Prague. Three groups of speakers reading the same sentences will be reported on here. The choice of students of Czech promised a certain homogeneity in age, previous education, interest in the study of their mother tongue, (partial) knowledge of the orthoepic norm and, last but not least, motivation. The groups of speakers can thus be described as representative of a higher level of pronunciation; as will be seen later, even here the number of deviations from the expected (orthoepic) norm is very high. It is obvious that these findings form a basis for appropriate (in some cases logopedic) measures and, hopefully, even for some changes in the curriculum of the Czech language. The first group of speakers in the first part of our investigation was formed by 33 students; the results are used here for comparison only. The remaining two groups, again students of Czech, consisted of one group of again 33 students, future teachers of Czech, whereas the additional group of 12 students was formed by students studying Czech without any qualification for a teaching job.

c) - an auditory analysis followed, performed (1) by a team of listeners in the first part of the project and (2) by a single listener, co-author of this paper, in the second part of our investigation; these results will form the core of our report. The previous results will be quoted for comparison only; some of them have been reported on at the Acoustic Conference in the High Tatra (October 1989). The task of the listeners was to transcribe the recorded text: in a preprinted form they had to write down all deviations from the expected orthoepic pronunciation. For the notation a code was used: 21 categories describing the quantitative and qualitative characteristics of speech segments. Some mispronunciations were expressed by a combination of the code "words": 22% of mispronounced vowels were described by more than one of the characteristics.

d) - results of the auditory analysis were then transferred to a database. The database formed then a starting point for a description of the actual pronunciation of our speakers, giving characteristics of speech of the whole group as well as data on individual speakers. Each DB record represented one speech segment (speech sound) deviating in some respect(s) from the norm as pronounced by one particular speaker. By a number of search routines and programs, the stored data were analysed from various points of view. To this end, the main file of deviations and the file containing detailed characteristics of the individual sounds in the text (initial-medial-final, vowel-consonant-syllable consonant, stressed - unstressed, member of a cluster) were joined, allowing thus a direct access to various categories of segments. The results of the searches were computed, transferred into tables and graphs, and interpreted.

Only some of the results can be presented here, giving information (1) about the performance of the speakers and their interpersonal variability and (2) about the degree of deformation of the individual speech sounds and the most common types of errors.

The attainments of the speakers are characterized by the number of mispronounced sounds (or by the total number of the errors which may be higher); deformations were found to form approx. 11% of the text (in our previous investigation in 1988: 20%). There are considerable differences between speakers: 8-33% errors (1988: 8-33%), 16% on the average (in the small group of 12 speakers: range 7 - 21%, average: 16% again.)

As for the types of mistakes:

1) of the possible 21 types of deformation, six types cover 90% (1988: 80%) of all deviations;

2) the most frequent deviation from the orthoepic norm is the extremely open pronunciation of vowels (though the speakers came from various parts of Bohemia and Moravia, not only from Prague and surroundings, where the open pronunciation is rather common);

3) next comes shortening (and reduction) of short vowels and shortening of long vowels, where, in the group of long vowels, it is the most frequent deviation;
4) an excessive nasalisation is the third
characteristic deviation. As for conso-
anents, weakening of articulation is here
the most common change.

The number of mispronounced vo-
wells is considerably higher than that of consonants: in 75% of the spe-
kers twice as much vowels are deformed
when compared with consonants. The
most common deviation is a too open
pronunciation, then shortening of long
and short vowels, reducing of vowel
quality, nasalisation, weakening of con-
sonants, omission and confusion of
sounds. Eight speakers in our sample had
a speech defect; in two other speakers the
nasality was excessive. Regarding the fre-
quence of errors in individual speech
sounds: more than 10% of errors were
found in consonants /f/, /l/, /s, /m, /v, /n/ and /l/
more than 15%), more than 5% also /s, /h, /z, /c.

In all, approx. 32 (1988: 36) % of all
vowels were deformed.

In short vowels the most frequent de-
viation is a too open pronunciation, then
comes a reduced timbre and changes in
quantity (both shortening and lengthen-
ing).

In long vowels an open pronunciation
and vowel shortening is very common.
The most frequent deviation in conso-
nants is their incomplete (weakened) re-
alisation; the speech defects are found in
sibillants and in /sound.

Perhaps some other findings may be
added:
- a fact, which may seem surprising
especially to speech therapists, is the
high number of mispronounced vowels as
compared with the consonants in the

The V/C ratio is 3:1 on the average,
i.e. generally there are three times more
mistakes in vowels than in consonants,
- some of the erroneous pronunciations
belong to the field of speech therapy
(though the number is not high and not
significant enough). Anyway, the number
of speakers with speech defects may
seem too high for future teachers of
Czech. A line had to be drawn, of course,
between occasional mispronunciations of a
"logopedic character" and real
speech defects. But even here the occa-
sional mispronunciations may point to a
certain instability in pronunciation;
- strangely enough, apart from the clear
"logopedic cases", the famous Czech /t
(Dvořák) remains unchanged.

A small table at the end of our paper
gives some general results, giving sums
and percentages of errors for individual
classes of speech sounds. Again, a con-
centration of deviations in the data for
vowels in comparison with those for con-
sonants is apparent here in somewhat
more detail. A correlation of these per-
centages with the results of the previous
parts of the analyses is high and significant
(r = 0.93).

Considerable differences can be seen
between the relative stability of the plos-
ives, a stronger tendency to deviations in
the group of fricatives and affricates and
the group of sonorants. Here again a
great difference between vowels and
consonants can be found.

These data are given here without re-
spect to the position of the speech sounds
within the text; as segments were coded,
however, with respect to their occur-
ance in initial, medial or final syllables,
in stressed or unstressed parts of the text
and also with respect to their positions
within clusters. This, of course, splits
the data into numerous minor groups. If
we tried to sum up simply some of these
results, then, in the first place, the fol-
lowing facts have to pointed out:
- differences in numbers of deviations
between initial, medial and final syllab-
es: not only final syllables show, as
could be expected, a higher number of
deviations, but also the sounds in initial
positions;
- no great differences were found in
results for stressed vs. unstressed syllab-
es.

In conclusion, two facts perhaps
derserve to be mentioned again: firstly,
a detailed analysis of our material reveals
a picture radically different from the sit-
uation with which speech therapists of te-
achers of foreign students are con-
fronted; secondly, the most common
and widely spread are those mistakes
originating in careless pronunciation
habits, leading then to reduced intelli-
gibility.

Numbers and percentages of
mispronunciations in

<table>
<thead>
<tr>
<th>Speech sounds</th>
<th>N</th>
<th>Err</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total:</td>
<td>14520</td>
<td>2473</td>
<td>17.0</td>
</tr>
<tr>
<td>Vowels (total):</td>
<td>5940</td>
<td>1899</td>
<td>31.9</td>
</tr>
<tr>
<td>Short:</td>
<td>4686</td>
<td>1591</td>
<td>33.9</td>
</tr>
<tr>
<td>Long:</td>
<td>1254</td>
<td>308</td>
<td>24.5</td>
</tr>
<tr>
<td>Consonants (total):</td>
<td>8560</td>
<td>574</td>
<td>6.6</td>
</tr>
<tr>
<td>Plosives:</td>
<td>2343</td>
<td>90</td>
<td>3.8</td>
</tr>
<tr>
<td>Fricatives:</td>
<td>2442</td>
<td>172</td>
<td>7.0</td>
</tr>
<tr>
<td>Affricates:</td>
<td>330</td>
<td>33</td>
<td>12.7</td>
</tr>
<tr>
<td>Nasals:</td>
<td>1551</td>
<td>74</td>
<td>4.7</td>
</tr>
<tr>
<td>Sonorants:</td>
<td>1914</td>
<td>205</td>
<td>8.3</td>
</tr>
</tbody>
</table>

N = number of sounds in a class
Err = number of mispronounced sounds
% = percentage of deviations (Err/N)

REFERENCES
FANT, G., L. HARD and A. KRUC-
KENBERG (1987) "Individual variations
in text reading. A data-base pilot study", RAUL 17, 194-114
JANOTA, P. and Z. PALKOVÁ (1989)
"Auditory analysis of speech segments" in Proceedings of the 29th acoustics con-
ference on physiology, acoustics, psy-
choacoustics, acoustic of music and of
speech, Štýrské Pleso, 194-197
PALKOVÁ, Z. (1990) "Stručné pravid-
al české ortoepie [Basic rules of Czech
orthoepy], Prague, Czechoslovak Radio
STRAATEGIES FOR PROSODIC PHRASING IN SWEDISH

Gösta Bruce*, Björn Granström** and David House**

*Department of Linguistics and Phonetics
Lund University, Helgonabacken 12, S-223 62 Lund, Sweden
**Dept of Speech Communication and Music Acoustics
Royal Institute of Technology, Box 70814, S-100 44 Stockholm, Sweden

ABSTRACT

This study focuses on the problem of prosodic phrasing in Swedish. A small database of sentences, potentially ambiguous with respect to phrase boundary location, have been recorded and analyzed. Considerable variation in phrase and clause boundary realizations was observed. Strategies including both boundary and coherence signaling have been identified.

1. INTRODUCTION

This contribution represents a cooperative work on a model for standard Swedish prosody in the context of a research project on prosodic phrasing in Swedish. The aim of the project is to investigate the phonetic correlates of phrasing using production data, text-to-speech synthesis and automatic prosodic recognition. In earlier work [2] we have outlined our joint research work on modeling Swedish prosody in a text-to-speech framework. See also [1] for earlier work aimed at developing a model for Swedish prosody, [3] for work directed towards the development of the prosodic component of a text-to-speech system, and [5] for a description of the prosodic parser.

It is widely recognized that grouping, involving the double aspect of coherence (connective) signaling and boundary (deaccentuating) signaling, is one of the main functions of prosody. Our focus of interest here is particularly in the division of an utterance into prosodic phrases and clauses.

The acoustic-phonetic signaling of prosodic phrasing is assumed to be complex, involving several parameters such as F0, duration, intensity, and voice quality as well as possible silence (physical pause). The more precise exploitation of these cues for prosodic phrasing in Swedish is, however, not well understood. The aim of the present paper is to explore different phrasing strategies which make use of some of these cues and their possible combinations.

2. SPEECH MATERIAL

In order to gain more knowledge about prosodic phrasing in Swedish [2], we devised speech material specifically designed for this purpose. As a starting point we chose sentences which, for the most part, were syntactically ambiguous. This was done to give us a preliminary idea about phrasing strategies and to enable us to easily test these strategies in the text-to-speech framework. The speech material consisted of 22 sentences, typically occurring as internal pairs, where the location of the sentence boundary was varied. Example sentence pairs are the following:

1a. Skolan börjar med samling i klassen. (School begins with a meeting of the class.)
1b. Skolan börjar, när barnen vågar. (School begins when the children dare.)
2a. När pappa fiskar, stör Piper Putte. (When daddy is fishing, Piper disturbs Putte.)
2b. När pappa fiskar, stör, Piper Putte. (When daddy is fishing, Piper disturbs Putte.)
3a. När han överlämnade sej och bonden hällade kungen med ett leende, så blev det bara så. (When he and the farmer surrendered, the king greeted them with a smile, that's the way it happened.)
3b. När han överlämnade sej och bonden, hällade kungen med ett leende, så blev det bara så. (When he and the farmer surrendered, the king greeted them with a smile, that's the way it happened.)

A male Stockholm Swedish informant read the speech material three times. He was given explicit instructions not to make any pauses at sentence/boundaries.

3. SPEECH ANALYSIS

In the present speech corpus, considerable variation exists in the acoustic-phonetic signaling of phrasing and phrase boundaries. Considerable variation in the realization of these cues and their possible combinations has been identified. In some sentences the FO of the sentence was observed. FO of the sentence we do not observe any obvious FO-boundary cues in connection with "fiskar", i.e. no FO-fall to a bottom level, although there are apparent segment lengthenings.

3.1. Boundary by duration only

One possible strategy is to use only duration for clause/phrase boundary signaling. This appears in some of the shorter sentences of our test material where there is no marking of the boundary in terms of FO. In these sentences we find segmental lengthening before the clause boundary. An example of this is given in Figure 1 where the final segments of the word "börjar" are clearly lengthened before the clause boundary (sentence 1b) as contrasted with the same word in the context before the prepositional phrase (sentence 1a).

3.2. Coherence by deaccentuation

Another strategy for prosodic grouping represented in our speech corpus is to use FO and duration (usually in combination) for the signaling of coherence within a speech unit. Exemplification is given with reference to the ambiguous pair of sentences 2a and 2b (Figure 2). In sentence 2b we observe the backgrounding of "fiskar" involving both flattening of FO deaccentuation) and segment shortening. The two words - "fiskar" (verb) and "stör" (object) - are produced as a unit with only "stör" being accented (focal accent). This unit accentuation serves as a connective signal and may by itself be sufficient for the disambiguation of sentences 2a and 2b. Usually, however, this coherence signaling is accompanied by explicit boundary signaling. A typical FO correlate here is the terminal FO-fall to a bottom FO level on "stör" (Figure 2b), which is also combined with segment lengthening.

3.3. Coherence by hat pattern

For the other member of the pair, test sentence 2a, with the intended internal boundary located between "fiskar" and "stör", we encounter another kind of coherence signaling without the use of deaccentuation. Here the FO rise on "stör" followed by the FO fall on "piper" together form a hat pattern [4], which serves as a connective signal. In this sentence we do not observe any obvious FO-boundary cues in connection with "fiskar", i.e. no FO-fall to a bottom level, although there are apparent segment lengthenings.

Figure 1. Partial spectrograms and FO of sentences 1a (top) and 1b (bottom).
3.4. Phrasing and syntax

Coherence signalling in the form of unit accentuation as exemplified here is restricted to certain syntactic constructions.

Sentences 3a and 3b are examples where deaccentuation does not apply. Here we find a more archetypical use of combine duration and F0 cues for prosodic grouping (see Figure 3). While the total duration of the two different readings (up to the final clause) appears to be the same, there are, as expected, local lengthenings at different places depending on the location of the internal boundary.

In test sentence 3b, where the boundary occurs after “bonden”, the pre-boundary lengthening is combined with a drop in F0 to a bottom level. This F0 drop is also the end of a typical downstepping pattern for the two last accents (“sej” and “bonden”) within the first phrase.

In the other member of the sentence pair, 3a, where the boundary is located after “sej”, we observe the pre-boundary lengthening as well as a fall to a fairly low F0 level, albeit not a bottom F0 level. When comparing the F0 valleys at “sej” and “bonden” across the boundary, there is no downstepping (pattern) to be observed.

The moderate F0 drop at “sej” in connection with the boundary in 3a, as compared with the drop to a bottom F0 level at “bonden” (where the boundary is in 3b), invites the following possible account of phrasing strategies. The syntactic structure of the two sentences displays an interesting difference. In 3a we have a coordination of two subordinate clauses before the main clause of the sentence, while in 3b a single subordinate clause precedes the main clause of the sentence, which is then followed by another independent clause.

According to our interpretation the moderate F0 drop at “sej” represents the sign of the continuation of the subordinate clause (in 3a), while the larger F0 drop at “bonden” (in 3b) represents the termination of this syntactic unit (subordinate clause).

4. CONCLUSIONS

We have identified and explored some alternative phrasing strategies in Swedish. Phrase boundaries can be signalled by duration only (pre-boundary lengthening) or by duration in combination with an F0 drop to a low level. Coherence within a unit can be marked by deaccentuation as well as by more complex means involving specific combinations of F0 and duration. Experiments using these strategies in synthetic speech and prosodic recognition will be reported at the congress.

ACKNOWLEDGMENT

This work has been supported in part by grants from the Swedish National Board for Technical Development and the Swedish Telecom. The present work is carried out within the Language Technology Programme under a grant from the Swedish Council for Research in the Humanities and Social Sciences.

REFERENCES

THE INTERACTION OF FUNDAMENTAL FREQUENCY AND INTENSITY IN THE PERCEPTION OF INTONATION

K. J. Kohler
Institut für Phonetik und digitale Sprachverarbeitung
Kiel, Germany

ABSTRACT

The temporal alignments of three terminal F0 peaks (early, medial, late) with stressed syllables, the parallelism of F0 and intensity timing in these patterns, and the importance of intensity in pitch accent signalling are discussed for German.

1. F0 PEAK POSITIONS IN TERMINAL INTONATION

In [4,5], I have shown that terminal intonation contours in German can have three different specific meaning related types of F0 peak positions around one and the same stressed vowel: (1) early, the peak may be early before the stressed vowel, which only gets an F0 fall (early peak); (2) the peak may be in the centre of the stressed vowel, which therefore has an F0 rise and an F0 fall (medial peak); (3) the peak may follow a stretch of low F0 in the stressed vowel and therefore not occur until the second half or even the beginning of a subsequent unstressed syllable (late peak). This means that the F0 rise dominates the stressed vowel and the F0 fall is not always realised in it.

The early peak differs categorically from medial and late ones by only having a falling F0 during the stressed vowel, thus accentuating the lower pitch range compared with the other two patterns. This categorical difference in the acoustic manifestation of early vs. non-early peaks is paralleled by a categorical change in perception along a peak position continuum from early to medial and by a continuous one from medial to late [3]. This means that for the signalling of an early versus a non-early peak a simple F0 fall as against the presence of an F0 rise is essential.

It follows from this that in the concatenation of F0 peaks without valleys between them ('hat patterns') [5], early peaks are not possible at the beginning of a hat, and non-early ones can only be signalled initially, if in the initial position of a hat the F0 fall is shifted further and further into the stressed vowel from an early via a medial to a late position, this shift lacks the change-over from fall to rise, because the preceding syllables are not lower in F0. Similarly, if in the initial position of a hat the F0 rise is shifted further and further to the right from a late via a medial to an early position, this shift lacks the change-over from rise to fall because the subsequent syllables do not have a dip in F0. In both cases we get continua of F0 fall and F0 rise, respectively, and the concomitant perception is equally continuous.

2. F0 AND INTENSITY TIMING

The precise F0 timing of terminal F0 peaks depends not only on the quantity and quality of the vowel, and the right-hand base point some 150 ms after the peak point. In early peaks, the peak point is preserved where medial peaks have their left-hand base point; the right-hand base point occurs at the end of a fall. When the F0 is sustained, the right-hand base point occurs at the end of a fall (short) or about 10 ms after the peak point.

In late peaks, the left-hand base point is positioned where medial peaks have their peak point, the stretch from the syllable beginning being low and descending slightly; the rise to the peak point then occurs within about 100 ms, after which we get a descent to the right-hand base point in another approx. 100 ms. This accommodates F0 time courses in late peaks the stressed vowels are lengthened after the left base point, more so for low than for tense vowels, more in final monosyllables than elsewhere. If voiceless consonants intervene between a lax stressed late peak vowel and a following unstressed syllable the target peak value cannot be reached in the stressed vowel itself, but is needed for pattern identification and therefore set at the voice onset of the following unstressed vowel.

In early and medial peaks, the low F0 fall at the end of an utterance is accompanied by a drop in source amplitude. In unstressed vowels and sonorants considerably, often reducing them to cracky voice and to irregular break-falloff. In early peaks this decline is shifted to the right following the later F0 fall, thus keeping a high source amplitude at the onset of unstressed vowels and syllabic sonorants; on the other hand the low F0 stretch in the stressed vowel before the peak gets its intensity reduced. So there is a natural parallelism in the time courses of F0, source amplitude and sound intensity for the three terminal peak contours. If it is destroyed in synthesis the output sounds either degraded or the peak pattern loses its identity.

The first case occurs, when a natural medial peak speech signal is taken as a point of departure for LPC resynthesis with a late peak in a completely voiced environment, as in 'She has b Ägelon.' (S he has been lying): the peak type is signalled correctly, but the utterance sounds...
husky at the end and overloaded in the middle because F0 and intensity are in opposite directions in these two places.

The loss of the particular characteristics of a peak pattern is illustrated by the synthesis of late peaks in an utterance-final word structure "stressed vowel + voiceless + syllabic nasal" as in 'Er ist ja geritten.' [...] (He has been riding.). A voiceless consonant after a late-peak stressed vowel interrupts the F0 course; it can only be successfully reconstructed by a listener if, in addition to an indication of a fast F0 rise speed (of ca 0.5 Hz/ms), the onset of voicing following the voiceless consonant receives the F0 peak and if the F0 descent from this value to the terminal low level can be clearly perceived. This means that the source amplitude must be high enough to guarantee sufficient intensity in the final nasal for the high falling F0 contour to be audible. If a natural medial peak speech signal with its low final intensity in the above utterance is taken for LPC reconstruction with a late peak, positioned at the nasal onset, the percept lacks the significant attributes of the late peak, because the intensity of the final nasal is too low and the F0 contour, therefore, not perceivable. Contrariwise, in a RULSYS TTS formant synthesis, by-rule of the above sentence [1], a reduction of the voice source A0 from 20 dB to 12 dB and of the nasal source from 30 dB to 10 dB in the final /n/ within a late peak (fig. 2) results in a loss of the perceptual late peak feature.

3. THE IMPORTANCE OF INTENSITY IN ACCENT SIGNALLING

The result shows that F0 and source amplitude are linked in production, and that their coupled time courses are expected by listeners. If the coupling is artificially destroyed by a listener there has to be sufficient voice intensity in the signal. In the examples discussed so far, an intensity reduction was capable of affecting the identity of a pitch accent, but not its presence, i.e. the stress position remained unaltered.

The question now arises as to whether it is possible to change stress perception simply by varying intensity. Obvious instances for testing this hypothesis are utterances that are ambiguous with regard to containing one or two stresses. If a late F0 rise is immediately followed by a medial F0 fall without an intervening F0 dip in two abutting stressed syllables, (fig. 1a), the second stress is weakened. If intensity alone can change stress perception, then it should be possible in a case like this to produce a switch in focus to initial sentence stress simply by reducing the intensity in the second accent and by simultaneously raising it in the first.

This has been interactively tested by changing the A0 values accordingly in the RULSYS TTS synthesis-by-rule. The result has been negative: the focussing and consequently the number of stresses, does not change; it is more the loudness relations that are affected. This is further support to the long-established finding that intensity has a low signalling value for stress compared with F0 and duration [2].

4. REFERENCES


Fig. 1: Phonetic transcription and F0 (squares and cosine interpolations), in the German sentence 'Der Ring glänzt.' (RULSYS TTS): a) two stresses: hat pattern, late rise + medial fall, b) two stresses: hat pattern, late rise + early fall, c) one stress: late peak. Horizontal: cs frames (cumulative and for each segment), vertical: Hz.

Fig. 2: Phonetic transcription, voice source A0 and nasal source AN (squares and cosine/2nd order interpolations) in the German sentence 'Er ist ja geritten.' with late peak (RULSYS TTS). Horizontal: cs frames (cumulative and for each segment), vertical: Hz for F0, dB for AO, AN.
RHYTHMIC PATTERNS OF THE DISCOURSE IN MEXICAN SPANISH AND BRAZILIAN PORTUGUESE

Antônio R. M. Simões

University of Kansas, Lawrence, USA.

ABSTRACT

The notion of syllabic foot is commonly used by investigators in determining the rhythmic patterns of languages, in terms of perception. Following this notion, Spanish and English are said to be, respectively, the typical cases of syllable-time and stress-time languages. It is very difficult, however, to confirm these rhythmic patterns empirically [7, 8, 9, 10, 11, 12, 15, 16, 23, 28]. Taking into consideration the recent discussions about P-centers, i.e. "perceptual centers" [7, 12, 16], an acoustical analysis was performed indicating that in Spanish, syllables may in fact have very similar temporal patterns, although Brazilian Portuguese (RP) may combine both the characteristics of syllable and stress-time.

1. INTRODUCTION

Linguistic studies have attempted to place natural languages into classes according to characteristic rhythmic patterns [3]. This notion is desirable because it has explanatory power for phonological processes in English, for example. Pike [26] explains that a reduction of the kind "If Tom will do it I will" (cf. "If Tom will do it will") may be explained if the notion of stress-time rhythm in English is used. And in fact, knowledge of the so-called "chopping" characteristic of the sentences in English is an enormous help to the foreign student in the classroom. In terms of the Spanish language, this author holds that the notion of vowel stability is more adequate than the notion of syllable-time. Syllable-time or staccato is a verbal impression and a consequence of vowel stability in Spanish. BP can be said to have both the stress-time characteristics similar to English and vowel stability depending on dialectical variation as well as intra-speaker variation. And this may be true of Spanish as well.

It may be that discussions concerning these notions are purely a matter of point of view. Although investigators suggest that BP has a stress-time rhythm, attempts to apply these perceptual notions to BP, not Peninsular Portuguese, seemingly have proved difficult as well [1, 2, 17, 20, 21, 27, 28, 29].

The notion of syllable and stress-time is a perceptual or impressionistic notion. Once we carry this notion to the physical measurements of syllables in sentences, the expected isochrony cannot be found. More recently the developments around the notion of P-centers [7, 9, 10, 12] may explain why subjects may have this perceptual knowledge of regularity although acoustically we find no correspondence. The regularity seems to be an underlying form which cannot be reflected acoustically. The works of Parker and Diehl [23] had already pointed out the possibility that the duration of a vowel may be greater than the acoustical signal tends to show.

The present study is a continuation of former investigations in the area of temporal patterns and their relation to rhythmic patterns. There will be no attempt to give a description of the structure of BP in this investigation for lack of space. Detailed and brief descriptive analyses of Portuguese and Spanish can be found in some of the works cited here [1, 5, 6, 13, 14, 17, 21, 22, 23, 24].

2. EXPERIMENTAL PROTOCL

The experimental protocol was organized according to three major procedures: the production of the recordings, the production of the spectrograms for sound segment segmentation, and data analysis. In the production of the recordings, passages from Mexican and Brazilian television broadcasts were recorded in the language laboratory at the University of Kansas by a laboratory technician. Recorded passages containing dialogues and news broadcasts were used randomly. Over one-hundred spectrograms were produced for analysis and measurement.

Segmentation procedures used in this study use Klatz's [18] way of segmenting, combined with the works of Leblanc and Peterson [19, 25] which deal with the notions of onsets,

offsets, steady state, and simple and complex nuclei, the work of Parker and Diehl [23], and the more recent notion of P-centers [7, 9, 10, 11, 12, 15, 16] as well. Detailed explanations as to the segmentation rules are given by the author elsewhere [27].

Two different methods of measurement were used. In the first method, only the vowel nucleus was measured, and in the second method, the vowel nucleus and the preceding consonant were measured when there was a preceding consonant. Otherwise only the vowel was measured. The statistical package SPS 4.1 for IBM VMCMS at the University of Kansas was used to run several different tests on segment's duration according to method, language, and relative position of the (consonant)-vowel to the stressed (consonant)-vowel. Before using parametric tests such as ANOVA, a comparison was done of the distribution of values using the median and the mean. Since no skewed distribution nor significant differences in values were observed, either the mean or the median could be used in this study. There were missing values in our data, but these were taken care of by techniques already existing inside the ANOVA program.

3. RESULTS AND DISCUSSION

The present results show for Mexican Spanish (MSP) a significant regularity of the temporal pattern of the sounds studied, regardless of the method. To the case of BP, different results will be obtained depending on the method used. Table 1 summarizes these results where MSP stands for "Mexican Spanish", BP for "Brazilian Portuguese", PR for "posttonic", and ST "stressed".

Table 1: ANOVA results of cell means and standard deviations by language, position and method.

<table>
<thead>
<tr>
<th>Language</th>
<th>Position</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSP (CV)</td>
<td>Mean 13.18</td>
<td>3.14</td>
<td></td>
</tr>
<tr>
<td>PR1</td>
<td>13.18</td>
<td>3.14</td>
<td></td>
</tr>
<tr>
<td>PR2</td>
<td>13.39</td>
<td>3.35</td>
<td></td>
</tr>
<tr>
<td>PR3</td>
<td>14.79</td>
<td>4.9</td>
<td></td>
</tr>
<tr>
<td>PR1</td>
<td>14.96</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td>ST</td>
<td>17.97</td>
<td>5.63</td>
<td></td>
</tr>
<tr>
<td>PST1</td>
<td>17.63</td>
<td>4.19</td>
<td></td>
</tr>
<tr>
<td>PST2</td>
<td>24.19</td>
<td>1.92</td>
<td></td>
</tr>
<tr>
<td>BP (CV)</td>
<td>Mean 12.37</td>
<td>2.96</td>
<td></td>
</tr>
<tr>
<td>PR1</td>
<td>12.37</td>
<td>2.96</td>
<td></td>
</tr>
<tr>
<td>PR2</td>
<td>13.64</td>
<td>5.95</td>
<td></td>
</tr>
<tr>
<td>PR1</td>
<td>14.35</td>
<td>4.47</td>
<td></td>
</tr>
<tr>
<td>ST</td>
<td>16.73</td>
<td>4.14</td>
<td></td>
</tr>
<tr>
<td>PST1</td>
<td>24.95</td>
<td>6.53</td>
<td></td>
</tr>
<tr>
<td>PST2</td>
<td>17.94</td>
<td>8.47</td>
<td></td>
</tr>
<tr>
<td>PST2</td>
<td>16.38</td>
<td>4.92</td>
<td></td>
</tr>
</tbody>
</table>

Preliminary analysis of the spectrograms containing samples of speech from MSP in this study have shown to be common for a vowel or a sequence of consonant and vowel in unstressed posttonic position to have longer duration than their stressed equivalents. This becomes even more evident when is treated as a pretonic position, confirming similar findings in what Klatz [18] called "pretonal lengthening". In the present study this syntactic and pretonal cut is not observed in BP which confirms results from an earlier study already undertaken [27].

This lengthening in MSP makes posttonic syllables longer than the stressed syllables in a discourse. This lengthening can also be observed by simply listening to a dialogue in Spanish in general, i.e. in any context. BP in this study confirms again results from Simões [27] done with the extreme vowels (i.e. where stressed vowels are twice as big as the unstressed vowel. The great posttonic reduction observed in that study was lessened in this study due perhaps to the great number of linking processes between words, more observable here. Pretonal lengthening, however, has not been observed here.

Other statistical tests were made, in an attempt to observe the relation between positions according to language and method as seen in Figure 1.
Figure 1: ANOVA results of multiple range test. The symbol * denotes pairs that are significantly different at the .05 level. Method-1

\[ M_{xy} \] (OY)

<table>
<thead>
<tr>
<th>BP (OY)</th>
<th>PR1</th>
<th>PR2</th>
<th>PR4</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.77</td>
<td>13.39</td>
<td>14.79</td>
<td>14.96</td>
</tr>
<tr>
<td>17.43</td>
<td>17.45</td>
<td>14.35</td>
<td>16.60</td>
</tr>
<tr>
<td>17.94</td>
<td>19.04</td>
<td>17.94</td>
<td>17.90</td>
</tr>
<tr>
<td>24.45</td>
<td>24.45</td>
<td>24.45</td>
<td>24.45</td>
</tr>
</tbody>
</table>

\[ M_{xy} \] (OY)

<table>
<thead>
<tr>
<th>BP (OY)</th>
<th>PR1</th>
<th>PR2</th>
<th>PR4</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.61</td>
<td>7.75</td>
<td>8.09</td>
<td>8.25</td>
</tr>
<tr>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
<td>9.00</td>
</tr>
<tr>
<td>10.48</td>
<td>10.48</td>
<td>10.48</td>
<td>10.48</td>
</tr>
</tbody>
</table>

Figure 1 suggests a much greater regularity in temporal patterns in MSp than in BP. The pairing of groups (syllables) as seen in Figure 1 indicates quite a different behavior in MSp. In BP the stressed syllable seems to be a function of reference similar to the stressed position in MSp. In the word "stop", strong positions in MSp are more evenly distributed among syllables, especially the stressed ones. This may be a result of syllables in general having relatively similar duration. Of course, the fact that prominence, stressed and positional differences are grouped in a way that is similar to the stressed position in MSp. In BP the stressed syllable seems to be a function of reference similar to the stressed position in MSp. In the word "stop", strong positions in MSp are more evenly distributed among syllables, especially the stressed ones. This may be a result of syllables in general having relatively similar duration. Of course, the fact that prominence, stressed and positional differences are grouped in a way that is similar to the stressed position in MSp.

The notion of prominence [17, 12, 16] has given the present analysis a clearer view of the temporal patterns observed. Although a perceptual analysis is necessary in this conclusion of this study, the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of the present results in Figure 1 from measurements at the acoustical level suggest that the nice model of measurements will provide a greater regularity in the temporal patterns of MSp.
SYNTAX AND INTONATION IN ITALIAN NOUN PHRASES

S. Kori and H. Yauda

Osaka University of Foreign Studies, Osaka, Japan.

ABSTRACT

The relationship between syntax and intonation in Italian noun phrases was studied. Acoustic examination of sentence-initial phrases in SVC sentences suggests that there are at least two syntactic factors that determine the tonal organization of an NP: branching construction and head-modifier relation. Branching construction triggers a boost of the protrusive F0 movement in the left-most content word in a constituent, and possibly an inhibition of the protrusion in other words. The head-modifier relation seems to cause a tonal fusion of two adjacent content words.

1. SPEECH MATERIAL

In order to examine the syntax-intonation relationship in Italian, F0 contours of six types of noun phrases were examined. The test phrases were put in a carrier SVC sentence 'NP è venuta/NP sono venuti a Padova' ('NP has/have come to Padova' (place name)). The internal syntactic structure of the test noun phrases were syntactically varied (Table 1).

Table 1. Test noun phrases. Content words are underlined and stressed syllables are italicized.

| 1. | [N Adj] & N | la doma brasiliana e il bismo |
| 2. | N & [Adj] N | il russo e il bravo brasiliano |
| 3. | [Adj] N & N | la bella brasiliana |
| 4. | Adj [N & N] | i giovani allievi e allieve |
| 5. | [N & N] Adj | la doma e il bismo brasiliani |
| 6. | N & [Adj] | la doma e il bismo brasiliano |

In all phrases, the number of inter-stressed syllables is kept constant at three. Three speakers from Northern Italy (SN and EF from Lombardia, and PC from Veneto) and a speaker from Central Italy (LT from Toscana) read the sentences four to eight times in a randomised order.

2. RESULTS AND DISCUSSION

F0 contours of the test noun phrases are plotted in the figures. The scale of frequency in the figures is logarithmic.

Some noun phrases have a different FO contour from others, suggesting that the internal syntactic structure is reflected in the tonal organization.

The syntactic difference between sentences 1 ([N Adj] & N: left-branching) and 2 (N & [Adj] N: right-branching) is realized in all speakers. In three of the four speakers this difference is realized in F0 contour. In sentence 2, the left-most content word in the right-branching constituent has a more conspicuous protrusive movement than in sentence 1: the left-most content word in the right-branching constituent has a conspicuous tonal protrusion.

In speaker RG, however, the difference in branching construction is reflected by the insertion of a pause.

A conspicuous tonal protrusion of the second content word is observed also in the right-branching construction in sentence 6 ([Adj N & N]), but it is less obvious in the left-branching construction ([Adj N] & N) in sentence 3. Thus sentences 2 and 6 have a more conspicuous protrusive movement in the second content word than sentences 1 and 3. In the former sentences, F0 contour in the second word is characterized by a rise followed by a fall, while in the latter sentences it is rather a break in the steep fall from the first content word, followed by another steep fall.

The conspicuous tonal protrusion due to right-branching, together with the conspicuous protrusion in the phrase-initial content word and the tonal inhibition of the phrase-final word, can be formulated as a general rule that the left-most content word in a branching constituent has a conspicuous tonal protrusion and other words inhibit their own protrusive movement.

However, the tonal protrusion due to right-branching in sentence 4 (Adj [N & N]) is observed in some utterances of speaker 4, but not in all speakers. Moreover, in some sentences with left-branchings, there is a conspicuous F0 protrusion in the second content word. In fact, the difference in branching construction between sentences 6 ([N & [Adj] N]) and sentence 5 ([N & N Adj]), which are a quasi minimal pair, is realized in none of the speakers because of a conspicuous protrusion in the second word.

The different tonal treatments for left-branching construction indicate that branching construction is not the only determining factor in the tonal organization of a noun phrase.

The syntactic difference between the sentence set 1 ([N Adj] & N) and 3 ([Adj] N & N) is the relation between the first two content words: in sentences 1 and 3, they are linked by a head-modifier relation, while in sentence 5 they are not linked by such a relation. This indicates that the local head-modifier relation is another syntactic factor determining phrase prosody: the second content word in the phrase which is not linked by a head-modifier relation with the first word has a conspicuous tonal protrusion in F0, whether it is the head or the modifier.

This rule predicts a more general rule that two content words linked by a head-modifier relation tonally fuse into one, inhibiting the protrusive movement of the second other words. The inter-subject inconsistency found in sentence 4 (Adj [N & N]) could be interpreted as an inter-word tone attack at the branching construction and the tonal fusion rule of the two words linked by a head-modifier relation.

3. CONCLUSION

Acoustic examination of F0 contours of the test noun phrases consisting of three content words suggests that there are at least two syntactic factors which determine the tonal organization of a noun phrase: branching construction and local head-modifier relation. Branching construction triggers a tonal boost at the left-most content word of a constituent, and possibly inhibits protrusive tonal movement of other words. Head-modifier relation appears to cause a tonal fusion of two adjacent content words, regardless of which is the head and which is the modifier, inhibiting the F0 protrusive movement.
of the second word, and thus its tonal independence. Two words not linked by such a relation do not tonally fuse. In cases where these two rules interfere, intra- and inter-speaker instabilities appear. The overall results lead us to believe that the syntax-intonation relationship in Italian is not linear in nature.

Figure 1. 
F0 contours of test noun phrases 
Speaker SG

Figure 2. 
F0 contours of test noun phrases 
Speaker EF

Figure 3. 
F0 contours of test noun phrases 
Speaker LT

Figure 4. 
F0 contours of test noun phrases 
Speaker PC
THE ROLE OF INTONATION AS A MARKER OF SEMANTIC ASSOCIATIONS AND ENUNCIATIVE OPERATIONS IN ENGLISH

J. Low

Laboratoire de phonétique, Département de Recherches Linguistiques, Université Paris VII. France.

ABSTRACT

The aim of this paper is to test the relation in English between the intonation of an utterance and the semantic content of its constituents. A corpus of utterances illustrating varying degrees of semantic associations read by several native speakers of British English was recorded. The analysis of the intonation contours shows small differences in fundamental frequency on the verbs of strong semantic associations and large differences in fundamental frequency on the verbs of weak semantic associations. The results are linked to enunciative relations and operations such as focalisation and modulation.

INTRODUCTION

The aim of this study is to test whether the intonation of an utterance is dependant or not on the semantic content of its constituents.

Many studies have shown the link between types of syntactic structures (Declarative statements, WH Questions and Yes/No Questions), parts of speech (content or function words), and intonation.

In order to isolate the problem of semantic content from that of syntactic structure and parts of speech the utterances studied were of the same syntactic type with the same number of content or function words.

CORPUS

The basis for this corpus was the work of Sheldon Rosenberg, the "Norms of Sequential Associative Dependencies in Active Declarative Sentences", in which he tested the link between the memorising ability of students on "semantically well integrated sentences" and "semantically poorly integrated sentences".

Two elements which are strongly linked semantically form a strong association and two elements which are weakly linked form a weak association. The type of structure for all the utterances in the corpus is: Noun Phrase (Determiner + Noun) + Verb + Noun Phrase (Det. + N). The subject (NP) is an animate noun, and the object (NP) an inanimate noun. The verb is in the present. Five basic sets of examples were chosen in which the noun phrases remained constant and the verbs expressed five varying degrees of semantic associations, e.g. for one set: constant elements The spider - the web, variable element: the verb, (1) spun, (2) made, (3) wove, (4) spoilt, (5) tore.

The basic sets are:
1. The actor - the part
2. The spider - the web
3. The author - the book
4. The priest - the sermon
5. The cat - the mouse

The 25 different utterances of the corpus were mixed with other utterances, and the order of the utterances illustrating the semantic associations was changed so that the informers were not aware of the aim of the test.

PROCEDURE

The material was presented individually to seven native speakers of Standard British English (3 women and 4 men between the ages of 22 and 26). They were asked first to read over the corpus thinking of the meaning of each sentence before recording. The recordings were listened to by 8 other native speakers who used the same phonetic system as those who produced the corpus. They were given generic samples of sentences to listen to and were asked, if they heard one word with greater prominence in each sentence, to mark that word.

An instrumental analysis was carried out on the recordings. The different contours were analysed according to measurements of fundamental frequency, time and the form of the end of the intonation contour.

RESULTS

The results of the perception tests show that the verbs which were part of a weak semantic association correspond to the point with the greatest prosodic prominence in the utterance whereas those that were part of a strong semantic association did not. The instrumental analysis shows the importance of the same phenomena: the prominent point within the intonation contour and the form or direction of the final part of the contour.

The contour was divided into a new segment at every change in direction. The different parts of the sentence were marked as follows:

Det Noun Verb Det Noun

The AB CDEF GHI JK LMN

In such a way, the segment GHI corresponding to the verb in each utterance of each set can correspond to a complex contour rise (GH) followed by a fall (HI).

A comparison of the differences in fundamental frequency (Fo) on the segments of the intonation contours in each utterance shows the following: small variations in Fo for verbs in strong semantic associations and large variations in Fo for verbs in weak semantic associations. A table showing the mean Fo differences for all the informers for the five verbs (1-5) representing different semantic associations in each set (I-V, 25 utterances) follows. Columns 2 to 5 represent the 5 degrees of semantic association, 1 being the strongest and 5, the weakest. The letters GH correspond to the rise and HI to the fall on the verb.

Table 1: Mean Fo differences on verbs (segments G-H, H-I) for 5 sets of utterances (I-V).

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>g - h</td>
<td>4</td>
<td>24</td>
<td>49</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>h - i</td>
<td>24</td>
<td>49</td>
<td>7</td>
<td>120</td>
</tr>
<tr>
<td>II.</td>
<td>g - h</td>
<td>12</td>
<td>7</td>
<td>23</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>h - i</td>
<td>25</td>
<td>27</td>
<td>35</td>
<td>103</td>
</tr>
<tr>
<td>III.</td>
<td>g - h</td>
<td>21</td>
<td>4</td>
<td>9</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>h - i</td>
<td>16</td>
<td>55</td>
<td>29</td>
<td>128</td>
</tr>
<tr>
<td>IV.</td>
<td>g - h</td>
<td>28</td>
<td>6</td>
<td>25</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>h - i</td>
<td>32</td>
<td>35</td>
<td>122</td>
<td>108</td>
</tr>
<tr>
<td>V.</td>
<td>g - h</td>
<td>5</td>
<td>41</td>
<td>31</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>h - i</td>
<td>39</td>
<td>5</td>
<td>80</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

Fig. 1 shows the mean Fo differences on the verb in set I.
Fig. 3a: The spider spins the web
AB CDEF GHI JK L MN

Fig. 3b: The spider toils the web
AB CDEF GHI JK L MN

Fig. 2 Mean total Fo for segments G-H, H-I for the 5 utterances of the 5 sets: 1x, 2x, 3x, 4x, 5x

The form or direction of the final part of the intonation contour varies, depending on whether the utterance corresponds to a strong or a weak semantic association.

In a strong semantic association the intonation pattern is a fall, and for weak semantic associations, the majority of the contours correspond to a final rise.

The intonation contours are shown in Fig. 3, illustrating a strong semantic association (a) and a weak semantic association (b), from set II produced by the same speaker.

In group 4 the subject (NP) the verb and the object (NP) correspond to notions with basic properties which are closely linked.

"A notion is a complex bundle of structured physico-cultural properties from which a notional domain is constructed with its formal properties such as the construction of a class and its linguistic complement" (A. Culioli). The relationships between the notional domains in the utterances in group A correspond to primitive relations, and the verb can not be focalised. Primitive relations depend on the notional status of the terms for they do not stem from any particular enunciative situation. A primitive relation is defined by A. Culioli as "a relationship between more than one notional domain, between the bundles of constituent properties which make up notions".

In group B, the notional domains corresponding to the verbs are not linked to those of the subject or the object. In this case the utterance can only be accepted if the verb undergoes an operation of focalisation marked by significant variations in intonation. The results in group A were produced with a final fall on the intonation contour and the majority of those in group B with a final rise.

The direction of the end of the intonation contour can be linked to the operation of focalisation.

Given a notion "P" topologically organized in an interior P ("what can be called P") and an exterior P ("what cannot be called P", or the linguistic complement of P) separated by a boundary F(P), the choice by the enunciator of either P or F is the modality of assertion (affirmative assertion for P, negative assertion for F). The inability to choose between P and F corresponds to the modality of interpolation.

The final fall corresponds to the choice of P or P (assertion). The final rise corresponds to the point in the operation of focalisation at which the choice between P and F cannot be made. Given this fact, it is interesting to note that, for the majority of the informers, the contours in group B correspond to a final rise. Thus, the validity of the assertion in that group seems to be questioned. What happens in fact is that, even though the utterances in group B are in the assertive modality, the weakness of the semantic link between the constituent notions generally makes it impossible for the enunciator to credit his own assertion with full validity. Therefore the interrogative intonation contour contradicts the assertive syntactic form.

The choice of the properties involved in the different notional domains represented by the predicate and the arguments in an utterance can thus be linked to the operations of focalisation and modalisation, as well as to the type of relation involved (either primitive or not).

This shows that neither syntax alone nor prosodic form alone can account for underlying operations. What has to be taken into account is the combination of the two kinds of markers.

REFERENCES


PERCEPTION OF INTOXATIONAL CHARACTERISTICS OF WH AND NON-WH QUESTIONS IN TOKYO JAPANESE

Kikuo Maekawa

National Language Research Institute, Tokyo, Japan.

ABSTRACT

The intonational difference between wh and non-wh questions in Tokyo Japanese was examined. Perception experiments involving synthetic intonation revealed that the most important cue for the discrimination between the two types is the lack of intonation boundary after the wh-word, rather than the prominence of the focused wh-word per se.

I. INTRODUCTION

That syntactic behavior of wh and non-wh questions differ is well recognized by grammarians. It seems to be less recognized by those who are working with Japanese prosody that the two question types differ significantly in their prosodic domains as well. As a matter of fact, the intonation does not consist in a mere difference of final rise but rather concerns the overall intonation shapes.

2. MATERIAL

Wh-questions are marked with wh-words like dare (who), doko (where), nan (what) etc. Incidentally, there are a class of words which are not wh-words but morphologically very similar to them: dareka (someone), doko (somewhere), nanika (something) etc. Those words are semantically marked, given their indefinite-pronoun-like meaning. As the result of their morphological similarity, we can construct pairs of wh and non-wh questions like (1) and (2), where syntactic and accentual configurations are exactly the same across two sentences. (Apostrophes denote accent locations.)

(1) na'ni-ga'yimai-e'ru [yp what-Now. 'see-Pot.-Pres.]
   = What can you see?

(2) na'nika'yim ai-e'ru [yp something see-Pot.-Pres.]
   = Can you see anything?

Fig. 1 shows typical examples of the F0 contours of (1) and (2) uttered by a male speaker of Tokyo Japanese (TJ). Their intonational difference can be examined in terms of their focus placement. Roughly speaking, the focus of a wh-question like (1) is on the wh-word, while the focus of a non-wh question like (2) is on its predicate. Usually the difference in focus placement is reflected in the prosodic structures of these sentences. According to the theory proposed by Pfeiffer-humbert & Beckman [1], the difference can be represented in terms of the difference of the 'intermediate phrase' defined as the domain of 'tactile sense.' While the whole utterance makes up an 'intermediate phrase' in (3), the utterance is divided into two different intermediate phrases in (4). It is interesting that the same prosodic difference can be observed in two 'accentless' Japanese dialects [2].

3. EXPERIMENT I

The aim of the first experiment was to examine if native speakers of TJ can in fact discriminate the two question types solely by means of intonation. The difference of (1) and (2) consists in the /k/-/g/ consonantal contrast as far as the segmental tier is concerned. So it was expected that subjects would be forced to rely on prosodic cues if we erased these consonants and then filled the resulting silence with white noise. On this reasoning, the following ten stimuli were prepared. The underlines show the time stretch replaced with noise.

(1a) naniga mi'eru
(1b) naniga mi'eru
(1c) nanigamieru
(1d) nanigamieru
(1e) nanigamieru
(2a) naniga mi'eru
(2b) nanigamieru
(2c) nanigamieru
(2d) nanigamieru
(2e) nanigamieru

In erase sequences of segments, care was taken to rid the effect of coarticulation as much as possible. Consequently, the white noise penetrates more or less into the final part of preceding segment and the beginning of following segment in all cases. All manipulation of original utterances, which were sampled in 10kHz/16bits

Fig. 1 The F0 contours of wh question /naniga mi'eru/ (left) and non-wh question /nanika mi'eru/ (right) as uttered by a male Tokyo Japanese speaker. The frequency scale is logarithmic.

Fig. 2 % correct identifications of wh question (real line) and non-wh question (dotted line). The abscissa represents the masking type indicated in the text.
condition. was made on a computer. These stimuli were presented to eleven speakers of TJ in random order in a quiet listening condition. The subjects were requested to identify whether the utterance they heard was (1) or (2). No notice concerning the relevance of prosody was given. Fig. 2 summarizes the result of the first experiment. Real and dotted lines show respectively the percentages of correct identification of wh and non-wh questions. The overall average correct identification rate is quite high (92.2% for wh's and 95.5% for non-wh's) showing that natural utterances are rich of prosodic cues. However, Fig. 2 provides us with little information about the relative importance of Pw and Sb. Both of these would seem to have equal importance in the identification task. (And it cannot be denied that cues other than the F0 shapes made certain contribution.)

4. EXPERIMENT 2

The aim of the second experiment was to examine the relative importance of Pw and Sb by using synthesized speech in which both cues were controlled. Fig. 3 shows the schematic structure of the stimuli synthesized. A-F of Fig. 3 denote the points where the contour is controlled. Point A is the beginning of the utterance, and is fixed at 200Hz. Point B is concerned with the cue Pw; its F0 value is either 300Hz or 250Hz. Point C stands for the beginning of the predicate niuru and is fixed at 140Hz. Point D is taken as representative of the cue Sb: the overall average correct identification rate is quite high (92.2% for wh's and 95.5% for non-wh's). Point E is the beginning of the sentence final rise and is either 150Hz or 200Hz. Point F is the target of the final rise and is fixed at 220Hz. For all the twenty combinations of the F0 values of B, D, and E, the four combinations in which the E value is higher than the D value were eliminated because these give rise to intonational configurations which are impossible in TJ. The remaining sixteen intonation contours were synthesized by the MACOR program using the PANAUTNS program developed by Hiroshi Imagawa and Shigeru Kirita-

The stimuli were presented to the same listeners in the same manner as in the previous experiment. Fig. 4 shows the percentages with which each stimulus was perceived as a wh-question. The abscissa of the figure is a composite representation of D values for the stimuli with E=120 Hz (the leftward three values) and for the stimuli with E=80Hz (the rest). The real and dotted lines stand respectively for the stimulus with B=300Hz and B=230Hz. This figure shows clearly that the contribution of the D value is greater by far than that of the B value. Although a raised B value (300Hz) makes some contribution to subjects' judgment of wh-questions, this effect is observed only when D is relatively high (180Hz or 160Hz). Once D is set to relatively low values (120Hz), the stimuli were perceived mostly as wh-questions irrespective of the B value.

5. DISCUSSION AND CONCLUSION

The two experiments reported here lead us to reconsider the phonetic nature of focus in TJ, stressing the importance of the salience of the prosodic boundary. In this respect, I believe that Fujisaki & Kawai[3] and Kori[4] have independently pointed out that focus not only increases the prominence of the focused constituent but also reduces the prominence of the following constituents. Kori also suggests that prominence of the final constituent of an utterance is more reduced than that of the other constituents. This analysis, which is based on production data, seems to be congruent with my perception data. Fig. 4 indicates that in order for a stimulus to be identified as a wh-question with 90% accuracy, it is necessary for the F0 value be lower than 120Hz i.e. lower than the right edge of the preceding NP. The data presented here and that of Kori and that of Fujisaki & Kawai suggest that any theory of phonetics that assumes that the effect of focus is limited only to the constituent marked as focused is inappropriate and to be revised. It should be pointed out that one important problem was left untouched: whether the difference of intonation examined in this study is specific to the pair of wh and non-wh questions. The line of reasoning that I followed in this study predicts that the difference is not a specific one. It is expected that the same intonational difference observed in any set of sentences having the same difference of focus placement as the one observed between (3) and (4).

6. REFERENCES


ACKNOWLEDGMENT

I am very grateful to Osamu Mizutani of NLRJ for his comments on an earlier draft of this paper.

204 205
COMBINATIONS OF TYPES OF PITCH ACCENT IN A CORPUS OF RUSSIAN SPEECH

Cecilia Odé
Institute for Perception Research, Eindhoven, The Netherlands

ABSTRACT

On the basis of a corpus of 15 minutes of spontaneous and prepared Russian speech, perceptually relevant pitch movements have been classified into types of pitch accent. A pitch accent is defined as a configuration of pitch movement(s) lending prominence to a syllable. The classification of pitch accents has been made by using the so-called stylization method (recently summarized in V. Hart, Collier and Corper (1990)). A number of perception experiments (Odé 1989) have resulted in 6 rising and 7 falling types of pitch accent. In the present paper combinations of types of pitch accent will be discussed.

1 PITCH ACCENTS

In tables 1 and 2 all types of rising and falling pitch accent, respectively, as observed in the corpus are given with their phonetic specification. The average values of all types of pitch accent are presented. Numbers between brackets indicate the maximum and minimum values of the features. These values are the limits of perceptual tolerance of the types of accent. The various types indicated in tables 1 and 2 are distinguished on the basis of the following features:

Direction distinguishes between rising and falling movements in the prominent syllable, that is between table 1 and table 2.

In the case of rising movements, excursus distinguishes between types R and r. Excursion indicates the size of an interval. In this article excursion is expressed in semitones measured from the lowest level of a speaker. For rising

there is a difference between a highest point reached within a range up to 10 semitones above the lowest level of a speaker (low register) and a highest point reached above the low register from 10 semitones up to the highest level of a speaker (high register). In the case of falling movements, excursion distinguishes between F and f. Timing indicates the position in the prominent syllable where the end frequency of a pitch movement is reached: the end frequency is reached near the vowel onset (early timing, symbol ‘−’), or much later than the vowel onset (late timing, symbol ‘+’). For rises, timing is relevant in combination with postonic parts (see below); for falls it is the only distinctive feature between accents F1-/Fnl− and F1+/Fnl+.

2 CONNECTING MOVEMENTS

Pitch accents are connected by non-prominence-lending pitch movements. These movements run from the (posttonic part of the) previous pitch accent to the (protonic part of the) next accent. The point at which a non-prominence-lending pitch movement turns from the last pitch accent into the non-prominence-lending pitch movement to the next accent, the so-called turning point (see the arrow in figure 1), is not arbitrary. Shifting the turning point forward or backward can affect the prosodic (and semantic) grouping of words. The location of the turning point is thus an important feature in non-prominence-lending pitch movements.

Figure 1: A turning point

3 PROSODIC BOUNDARIES

Table 3 gives all sequences of two successive types of pitch accent between prosodic boundaries that were found in the corpus.

The perception of a prosodic boundary is cued by pitch and/or temporal organization of an utterance. Prosodic boundaries (...) are perceived as clear breaks in the speech stream although acoustically silent pauses need not be present (J.J. de Rooij 1979:143). A prosodic boundary is relevant for the semantic organization of an utterance. The position of the boundary can mark the end and beginning of a stream of thoughts. Generally speaking, a prosodic boundary is heard as a pause within or at the end of an utterance. Prosodic boundaries were also perceived at a silence, a hesitation, a reset (an abrupt jump upward or downward in the F0 course) and at a turning point between two pitch accents.

In spontaneous speech elliptic phrases frequently occur. However, sudden interruptions in an utterance do not always correspond with interruptions or F0 changes in the melodic course of an utterance. In the corpus I have marked prosodic boundaries at positions where clear breaks in the speech stream were perceived. My observations have been verified by two highly trained listeners, native speakers of Russian.

4 COMBINATIONS

A combination of pitch accents is a sequence of pitch accents between prosodic boundaries.

Types of pitch accent that usually occur as the last accent before a boundary are types Rl−, Fl−, Fnl−, Fl+−, Fnl+. Types Rh− and R− regularly occur both as a last accent before a boundary and as a non-last accent. I will now discuss the single examples of sequences where these accents do not occur before a boundary. The numbers between brackets after the examples refer to pages in Odé 1989. The type of pitch accent is indicated directly after the word in which it occurs.

Type Rl−, if not before a boundary, can be followed by types Rm−/+/ and rm−/+. An example is to nabor (Rl−) kako (rm−/) (267), where the two pitch accents immediately follow each other. The same phenomenon was observed in other cases. Type Rl− followed by type Fnl− is observed in the utterance nam nayu poeche (Rl−) vot na ju (Fnl−) (230); and type Rl− followed by type Fl− is observed in the utterance in ty o ty o belostaja vasilj (Rl−) ne chvatilo (Fl−) (252). Type Rl− is followed by Fl− in nu pometa (Rl−) ruketa (Fl−) (254). In all these cases there is a direct connection, semantically and syntactically, between the two pitch-accented words.

Type Rl− can be replaced by types Rm−/+/ or rm−/+, but that accent is less emphatic.

Type Rh−, if not before a boundary, can be followed by the same accent or by type Rl−/+/, Fl−/−, or Rl−/− before the utterance is completed before a boundary with the accents Fl−, Fl+/− or Fnl+. I think it is just by chance
that type Fn+ did not occur after type Rh— in the corpus. In an experiment (Ode 1960:61-64) it has been established, that type Rh— is soon followed by a final fall, and only occasionally are some accents realized between type Rh— and the final fall. For example: ja repetirovao (.-.) veznu (Rh—) Sentak (Rm—) i ego sestu Ljubovu (F1+) (213); no podvizi (Rh—) poka (Rm+) mikroskopija (F2+) (265).

In contrast to Nikolaeva's findings (1974-4), in my material there is no phonetic difference between types Ri—/Rh— in a final clause of a sentence and in non-final clauses. Both types occur in both positions, with different sizes of excursions, but the excursion is always large.

Type Rm—, which is frequently followed by a final fall, is in one case followed by type Fm+ in the exclamation čert (Rm—) ego zvet (Rh+) (282). An example of Rm— followed by type Rm+ is: gm oni (Rm—) estradni (Rm+) Akametni nauk (231).

The final falls Fi—, Fn+, Fi+ and Fn+ have been found after one another, for example afterthoughts: v pijat (Rn—) pijat desjat (Fi—) oituda (Fi+) (249) or in the final sentence nal'ostal se shrotio na čevu (Fi+) raket (Fi+) (252). It is interesting to see that most of the sequences of final falls within one utterance occur in the most lively dialogue of the corpus. Other examples of sequences of falling types of pitch accent are: osnovok (Fi+) ovet tak olovko (Fn+) (231); na čevu Kolja Grmenczky (Fn+) akzal (Fn+) (284); osnovok ovet granov (Fm+) topli a (Fi+) (231).

Types Fn— and Fn+ are followed by the rises Fi—, Rh— and Rm+ in a few cases. Probably because of the high speaking rate in the spontaneous fragment no boundary was perceived after the fall. Examples are: cuti to akzal (Fn+) perestrojka (Ri—) (256); v tri časa idet klataja (Fn+) raket (Ri+) (256); primo skazan normal'mo (Fn+) suszpadite (Rm+) temperatury (Rh—) (256); muzit propitki (Fn+) tuda (Fn+ overs) um (Rm+) (256).

Finally, type Fh— can be followed by the same type: ona (...) točn (Fh—) sokstvotvornu (Fh+) (219) and by type Fh+; da eto (Fh+) dita menje v odljem ćitom 'nica sud' ('kresten'nyi) (Fn+—) kor- pros (Fn+) (233).

Type F+ is a repetition of the same pitch accent (see table 2) and will not be discussed here.

5 TOWARDS A LINGUISTIC INTERPRETATION

A type of pitch accent can have various functions in different contexts; different types of pitch accent can be used in one function. In my opinion, for all examples of one type of pitch accent in the corpus, the contextual functions of that type should be examined in order to determine whether contextual functions can be summarised into one meaning. If that is the case, the contextual functions found are interpretations of that meaning. Realisations of one type of pitch accent are perceptually equivocable, but contextual functions differ.

For example, type Fh— is interpreted as a question in the affirmative (Fh—) (231). In the utterance a osnovno i vozmno type Fh— is interpreted as the punctuation mark '!' in the utterance osnovno i vozmno (Rm+). Sspon manu (Fh—) v takom dučke (Fh—), the stream of thoughts is incomplete and evokes a question. On the other hand, in questions and incomplete utterances we also find type Fh+, e.g. in one example atendency (Fh+) and oni nekolu oituda (Fh+). At the congress more examples of combinations of pitch accent will be presented with their interpretation.

6 REFERENCES


Table 1. Types of rising pitch accent: average values and minimum and maximum values (limits of perceptual tolerance). R = rise with large excursion, r = rise with normal excursion, l = low, p = postonic part, n = no postonic part, m = middle postonic part, e = early timing, w = late timing.

<table>
<thead>
<tr>
<th>Type</th>
<th>Excursion</th>
<th>Timing</th>
<th>Postonic</th>
<th>Slope</th>
<th>Repeat</th>
<th>Picture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rm+</td>
<td>17 ST</td>
<td>(13.21)</td>
<td>60% early</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>R+</td>
<td>17 ST</td>
<td>(13.21)</td>
<td>60% late</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>R-</td>
<td>16 ST</td>
<td>(12.21)</td>
<td>60% late</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Rm+/r</td>
<td>15 ST</td>
<td>(11.17)</td>
<td>60% late</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Rm+/l</td>
<td>15 ST</td>
<td>(11.17)</td>
<td>60% late</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Rm+/r</td>
<td>13 ST</td>
<td>(11.15)</td>
<td>60% late</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Rm+/l</td>
<td>13 ST</td>
<td>(11.15)</td>
<td>60% late</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
</tbody>
</table>

Table 2. Types of falling pitch accent: average values and minimum and maximum values (limits of perceptual tolerance). F = fall, f = fall to a level above non-low; the lowest level is not reached, h = high postonic part, f = fall to a level above non-low, m = the configuration is repeated, w = late timing.

<table>
<thead>
<tr>
<th>Type</th>
<th>Excursion</th>
<th>Timing</th>
<th>Postonic</th>
<th>Slope</th>
<th>Repeat</th>
<th>Picture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
<tr>
<td>Fm+</td>
<td>8 ST</td>
<td>(6-11)</td>
<td>42% low</td>
<td>low</td>
<td>85.2%</td>
<td>(44.12)</td>
</tr>
</tbody>
</table>

Table 3. Sequences of types of pitch accent: the sign x indicates which type of frequently occurring pitch accent can be followed by which other type of pitch accent in the corpus. The pitch accent Rm+/r, r/r, r/r and f/l all occur with early and late timing. Types Rm— and Rm+, etc., are not discriminated on the basis of early or late timing. Therefore, the intonation '-'/'-' has been left out of this table. Single cases are indicated with the sign 0.

<table>
<thead>
<tr>
<th>Type</th>
<th>Rh-</th>
<th>Rm-</th>
<th>Rn-</th>
<th>Rh+</th>
<th>Rm+</th>
<th>Rn+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rm+</td>
<td>0</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Rm-</td>
<td>0</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Rn-</td>
<td>0</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Rh-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

208

209
A CROSSLINGUISTIC DESCRIPTION OF INTONATION CONTOURS OF A MULTILANGUAGE TEXT-TO-SPEECH SYSTEM

G. Olázy
Phonetics Laboratory, Linguistics Institute of the Hungarian Academy of Sciences, Hungary

ABSTRACT
Building elements to realise intonation contours in the MULTIVOX multilingual text-to-speech system are discussed. The description concerns intonation patterns on the word, phrase and sentence levels from the point of view of Hungarian, German, Finnish, Italian, Esperanto and Esperanto. The crosslinguistic features of the patterns will be shown as well.

1. INTRODUCTION
In text-to-speech synthesis the robot-like sound can be improved towards a more natural, human-like voice quality among other things by superimposing intonation patterns. The newest directions in text-to-speech synthesis point in many cases towards a multilingual approach combined into one modular system [2]. The MULTIVOX system is a general, text-to-speech system developed in Hungary [4] for multilingual synthesis. The system works in Hungarian, Italian, German, Esperanto and Finnish. New languages can be adapted easily to the basic system. Dutch and Spanish are under development. The synthesis hardware is the PCFS2000 format synthesizer. In MULTIVOX a modular representation of intonation has been implemented.

2. ELEMENTS FOR INTONATION AND STRESS
In devising acceptable intonation for unrestrained text we must formulate a set of rules which result in natural sounding pitch contours for utterances that may have never been spoken [9]. In the MULTIVOX system the following elements of pitch movements and timing contours are used as modular units in intonation and stress generation.

1. Starting (S) point of the pitch contour
2. Direction of the pitch movement: rise (R), fall (F)
3. Degree: high (H), medium (M), low (L)
4. Steepness (St) of movement in time
5. Jump down: (Jd) + (level) or (S) or (E)
6. Jump up: (Ju) + (level) or (S) or (E)
7. No change (N)/
8. End point (E) of the pitch contour
9. Lengthening (L) of the stressed vowel

The parameter degree can be adjusted to all units. Examples: RM means rising to medium level; Ju(SM) means jump up to a medium starting point. The physical values concerning these three degrees are shown in Table 1.

<table>
<thead>
<tr>
<th>Unit</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>S/E</td>
<td>25</td>
<td>10</td>
<td>95</td>
</tr>
<tr>
<td>R/F</td>
<td>25</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>St</td>
<td>2</td>
<td>0.5</td>
<td>2x</td>
</tr>
<tr>
<td>L</td>
<td>3x</td>
<td>1.5x</td>
<td></td>
</tr>
</tbody>
</table>

These data are used for a male voice generation.

3. PITCH AND TIMING IN WORD STRESS
Two questions were taken into consideration in the formation of word stress, the relation of pitch variation and the duration of the vowel in question.
(i) Whether pitch change occurs with lengthening or not? This and the place of the accent is shown in Table 2. Table 2 expresses that in Italian and Esperanto the pitch contours and the lengthening of the vowel in question have to be treated together. In the other languages these two parameters are treated separately.

(ii) Vowel duration influences the form of the pitch pattern. Our experience is that the same pitch contour cannot be used automatically in the case of a short and a lengthened vowel. Slight changes characterise the pattern for long vowels. A stress pitch contour for these cases looks like this:

For a short vowel (V):
SM(RM)(SH)+SM(EM)

For a long vowel (V):
SM(RM)(SH)+Nx+SM(EM)

The value of (x) is language dependent. For Hungarian and German it is cca. 30 ms, for Italian and Esperanto in closed syllables cca. 30 ms, in open syllables cca. 60 ms, in Finnish cca. 60 ms.

3.1. Four cases:
Rule 1: Stress on the first syllable. Languages: Hungarian, Finnish, German
Rule 2: Stress on the last syllable. Languages: Italian, German
Rule 3: Stress on the penultimate syllable. Languages: Italian, German, Esperanto
Rule 4: Stress on other syllables. Languages: Italian, German
Rule 5: Unstress the sequence. Languages: all.

These rules are based on word stress realisation in the mentioned five languages.

3.2. Algorithms for stress assignment
As Table 2 shows, Hungarian, Finnish and Esperanto can be treated as fixed stress languages, German and Italian are free stressed ones. For fixed stress languages the stressed syllable in the word is determined by the rules 1 and 3. If the stress is signalled by disc-ritics – like in Italian –, rule 2 will be used.

For free stress languages the algorithms for finding the stressed syllable in the word are based in many cases on a large morpheme inventory (10,000–50,000 entries) and a morpheme analyser algorithm. Such solutions are known for English [1] and for German [3] and for Italian [7], too.

The MULTIVOX system was designed to work with a relatively small memory (max. 100 kbyte) and in real time on a PC. Therefore no morpheme inventory and no morpheme analysis is used at all. To assign the proper place of the stress in the word (for Italian and for German) the "letter sequence" method (LSM) [5] and some other special algorithms were developed. The output of LSM is a sound level representation of the written text where the final duration of vowels is already set correctly in 95% (incorporating the necessary lengthenings coming from stress or from other linguistic rules).

In the Italian version of MULTIVOX the stress algorithm searches the syllable to be stressed on the basis of vowel durations. The stress will be superimposed where a vowel is lengthened in the word. This solution is an indirect approach to stress assignment. A more complicated solution appears in the German version, where the place of stress was assigned by the following rules:

D1. There is only one stress in one word.
D2. Stressed prefix suffix has priority against other rules (ankommen, Kom- men, etc).
D3. An unstressed prefix is followed by a stressed syllable (bekommen, gesagt).
D4. In two syllable words the long vowel if any is stressed (Sagen, Suchen, primär), else the first (Sile, Tsunten). This last rule is based on empirical observations.

Using these rules for finding the place of stress in German words a correct pitch superimposition is performed in 95% of the cases. The evaluation of these rules were done by listening to 1600 German sentences [8] and 20 text files (one A4 page each) gathered from books and newspapers. A weaker point of the German word stress assignment is the cases of compound forms. Here only rules D2 and D3 can assign a place of the stress for pitch patterns. Incidentally, the correct timing structure (without a pitch pattern superimposed) gives the feeling of correct stressing in most cases.
3.3. Pitch patterns for word stress

The following types of pitch patterns (PP) are used to create the frequency component of stress:

PP1.SL+RM(SM)+FH(SH)+EM

Hungarian: first syllable, Italian: every stress except final, German: stressed suffix

Esperanto: every stress.

PP2.SL+RM(SM)+FH(SH)+N(x)+Ju[EM]

German: first syllable in more-than-two-syllable words.

PP3.SL+RM(SM)+FH(SH)+N(x)+Ju[EM]

Finnish: every stress.

PP4.SL+Ju(SM)+PP2

German: unstressed prefix in more-than-two-syllable words.

PP5.SL+Ju(SM)+PP3

German: unstressed prefix in two-syllable words.

The question of unstressing is just as important as stress if we want to get close to the natural variation among stressed, unstressed, and neutral words in human speech. Unstressing in MULTIVOX is generated by reducing the pitch value to SL during the sequence (word, prefix, suffix etc.). This method is used in all the languages in the system. In sum, concerning word stress generation three types of cases are used: stressed, unstressed and neutral sequences. All these sequences remain present in higher level intonation patterns, i.e. in phrase and in sentence intonation.

4. PHRASE LEVEL INTONATION

The detection of phrase boundaries is performed in general on the basis of parsing [1], [3]. The MULTIVOX system is irregular with respect to this solution too. A simple phrase boundary detection was designed and realised, similar to the solution proposed by O'Shaughnessy [6] for English. Function words and some other special words are used to detect boundaries [5]. This solution is done for all the languages. Exceptions are Esperanto and German, where additional rules also help to improve phrase detection. For Esperanto noun and verb phrases can be detected because of the regularity of the language. In German the nouns are detected by searching capital letters as initials in words. For phrase intonation the same pattern is used in all languages i.e. the pitch is slightly raised continuously in the last two syllables of the phrase e.g. Ri(SM). The pitch is set back (JuM) during the phrase cycle which is 200-300 ms between the phrases.

5. SENTENCE INTONATION

In sentence intonation a serious problem is to find such rules that make the monotonous sounding more natural, so that listeners to long texts should not be uncomfortable [2].

Two types of sentence intonations are generated automatically in the MULTIVOX system, one for declaratives and one for questions. For declaratives the general theoretical pattern is a linear falling one. This pattern is used in all the languages except Italian, where a rising-falling pattern is superimposed. To achieve variability in long texts (sentence by sentence) the following simple rules were built into declarative intonation: the starting pitch value and the steepness of the declination is changed as a function of sentence length (Table 3).

<table>
<thead>
<tr>
<th>Sentence length</th>
<th>Start pitch Steepness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short (300ms)</td>
<td>120 Hz/100ms</td>
</tr>
<tr>
<td>Medium (600ms)</td>
<td>116 3</td>
</tr>
<tr>
<td>Long (900ms)</td>
<td>114 0.5</td>
</tr>
<tr>
<td>Very long</td>
<td>112 0.2</td>
</tr>
</tbody>
</table>

In addition, the last word of the sentence is set to a lower pitch value for creating the feeling that the sentence has ended. At the phrase boundaries the pitch is set higher (1-2 Hz/boundary) in the long and very long categories. This gives the feeling that a new phrase has begun. With these simple rules a relatively diversified sounding has been reached in reading long texts.

In questions, different types of pitch patterns have to be superimposed depending on the kind of question, like question with Q word/without Q word, one-syllable question.

5.1. Question with Q word

A general regularity is used for all the languages in the system. A high peak is set on the Q word i.e.

RH(SH)+FH(SM)+FL(SM)

and the one-falling pattern is superimposed (similar to the declarative sentence but with less steepness). It is important to set the end of the falling part of the peak lower than the starting point of the peak. The place of pitch change depends on the Q word end on the language (first second etc syllable). Markers sign the subgroups of Q words and the peak is placed where the marker points.

5.2. Questions without Q word

A general pattern for all the languages – except Hungarian – is as follows: The beginning is Jd(M) and the end is like in the phrase pattern. It is important to set a lower starting point than in the declarative sentences. In Hungarian the end pitch value is also like. The following patterns are thus used: RH(SH)+FH(SM) on the penultimate syllable.

5.3. One-syllable questions

The pattern is the same for all the languages except German. In question patterns, this is a rising one i.e.

SL+RL(SL)+RL(SM)+RH(SH).

This pattern expresses a gradually increasing pitch value in the question.

6. CONCLUSIONS

An attempt at multilingual intonation synthesis with a limited number and sort of pitch patterns was described. Our findings are that the patterns shown above are enough to realise the most characteristic pitch contours of many languages. The practical working of the above patterns was tested in the MULTIVOX system. The results are tolerably good.

7. REFERENCES

MEASURING INTONATION AT LOW SIGNAL-TO-NOISE RATIOS

V. Piktuna

Technological University, Kaunas, Lithuania

ABSTRACT

The method is proposed for evaluating intonation curve from the highly corrupted speech signal. During local processing the adaptive threshold is applied to the short-time FFT-spectrum, pitch harmonics are identified and pitch frequency determined. During global processing the intonation curve is smoothed and approximated by the low-order polynomial.

1. INTRODUCTION

Evaluating intonation when signal is corrupted with noise is a problem of great difficulty, especially in speech communication systems where only the past of the signal’s properties can be taken into consideration. There are however applications where measuring in real time is not necessary, e.g. teaching of deaf persons to speak, studying foreign languages, speech rehabilitation after operations etc. In these cases, uttering must be followed by an intonation curve on the screen for visual comparison to a reference one. This situation is less complicated because of the intonation contour possible, and both past and future values can be taken into account at every point of it.

When measuring intonation from spectral data, identifying of pitch harmonics simplifies calculating of pitch frequency (PF). The method is trended towards looking for periodicity in the corrupted spectra of speech, so it can find a “pitch” in the spectra of noise too [2]. Therefore the great attention is payed to recognition of noisy frames. The essential features of the method proposed are:

1. employing of the adaptive threshold (ATH);
2. identifying of pitch harmonics by their amplitudes, shapes and symmetry;
3. usage of a multistage procedure for the voiced/unvoiced decision.

The block diagram of the algorithm is presented in Fig.1.

2. IDENTIFYING OF HARMONICS

2.1. Evaluating of the Short-Time Spectrum

We suppose at least three pitch harmonics to be necessary for taking decision about the PF. If the highest PF for a female speaker is 450 Hz then the frequency of the harmonic under consideration must be at least 1350 Hz (1430 Hz in our hardware). The signal is weighted by the Hamming window and zeroes are added to obtain the FFT spectrum (in the logarithmic scale)

At 64 spectral points. The spectral resolution is 22.3 Hz, the measuring accuracy is improved by parabolic interpolation of spectral peaks.

2.2. Adaptive Threshold

A horizontal threshold has a primary disadvantage related to the formant structure of the spectrum: it can either not reach harmonics in the region between formants or cross the spectral components related to background noise. The ATH is obviously necessary changing when the spectral properties of the speech signal change. We propose for this purpose the spectrum of the linear prediction (LP) model. As the narrow frequency band is considered, the low-order LP models can be used. Fig.2 illustrates the effect of thresholding for different sounds and signal-to-noise ratios (SNR), when the ATH is of the type:

$$\text{ATH} = 20\log(1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3})^2$$

$$\sigma(\tau)$$ being the LP coefficients, $$z = e^{2\pi i \omega}$$, $$\omega$$ being the current frequency.

The value of shifting downward the ATH depends on the SNR and is discussed in [2].

2.3. Examination of Spectral Peaks

The three parameters of every spectral peak exceeding the ATH are examined: amplitude, sharpness and symmetry. The amplitudes are calculated directly from the spectrum (see e.g. [4]) while sharpness and symmetry are evaluated by the parabolic approximation of a spectral peak:

The coefficient $$\alpha$$ of a parabola and the approximation error correspondingly. The ranges of values for these parameters are defined in advance, using statistics of natural speech [2]. A spectral peak is considered a pitch harmonic provided all the three parameters are within the ranges defined.

3. CALCULATING OF THE PITCH FREQUENCY

The data for calculating PF are $$f(\omega)$$, the frequencies and $$\alpha(\omega)$$, the levels of maxima of spectral peaks. Obviously, $$\alpha$$ is not a number of a pitch harmonic. We have chosen a method of evaluating PF carrying the visual one: we consider the average distance among harmonics to be the PF. The evaluating is carried out in 2 steps:

1. the initial value of PF is calculated as the average distance among three harmonics: one of the maximum $$\alpha(\omega)$$ all over the spectrum and two closest to it (one from the left and another from the right). The possibility of lacking one (or two) harmonics among these 3 ones is accounted. Such an approach allows to find a curve of the PF even of high corrupted signal. We find this approach more reliable than those concerning spectral peaks starting from the very first on the left (e.g. [4]), if no equidistance among the three harmonics can be found, the same procedure is repeated with the other three ones in the neighbourhood (on the left and, if necessary, on the right).

2. the distances between all harmonics approximately equal to the initial value are averaged.

4. RECOGNITION OF UNVOICED FRAMES

4.1. Spectral energy

The unvoiced sounds are of little low-frequency energy.
We have empirically fixed the level of \(-10\ldots15\) dB for a horizontal threshold which must not be exceeded to identify the corresponding frame as voiced (Fig.1, V/UV1). This scheme works reliably at high SNR only.

4.2. Smoothing and approximating

The 3-points nonlinear smoother [3] and polynomial approximation are applied to the intonation curve. When approximating by a polynomial, the question arises how long must be the segments under approximation. Approximating of every voiced segment and of the whole curve are two extremities. Fig.3 shows the intonation curve consisting of 5 voiced segments where 3 and 2 segments are approximated by the 3rd and 4th order polynomials.

6. RESULTS

The method was tested with 3 speakers (two males and one female) using a limited speech material. When using knowledge of a human expert, the intonation curve remains at SNR down to 0 dB.

7. REFERENCES


SPEECH F0 EXTRACTION BASED ON LICKLIDER'S PITCH PERCEPTION MODEL

Alain de Cheveigné
Laboratoire de Linguistique Formelle, CNRS - Université Paris 7, France.

ABSTRACT
According to a pitch perception model proposed by Licklider [1, 2, 3], time-domain patterns of activity in nerve channels coming from the cochlea undergo autocorrelation analysis in the auditory nervous system. We examine whether this model can be adapted to the task of speech F0 estimation, and in particular what benefit the filter-bank processing stage can bring to a fundamental period estimation algorithm. Results show improvement in reliability over the same algorithm applied directly to the speech signal.

1. INTRODUCTION

1.1. Perception models applied to F0 extraction
A large number of F0 estimation algorithms have been proposed [4]. Some are purely signal processing methods, others derive from models of speech production or perception. While they mostly give similar results on clearly periodic voiced speech, some may fail or give doubtful results on less periodic or unvoiced speech. The result of their processing is a measure of neural activity over the dimensions of frequency (inherited from cochlear filtering) and lag (implemented as nerve conduction or synaptic delay). In response to a periodic stimulus such as voiced speech, a ridge appears spanning frequency at a lag equal to the period. The position of this ridge is the cue to pitch. Licklider's ideas have been developed recently by other authors [12, 13, 14, 15, 16].

1.2. Licklider's model of pitch perception
Licklider [1, 2, 3] proposed a model according to which each channel within the auditory nerve is processed by an autocorrelation mechanism. The result of this processing is a pattern of neural activity over the dimensions of frequency (inherited from cochlear filtering) and lag (implemented as nerve conduction or synaptic delay). In response to a periodic stimulus such as voiced speech, a ridge appears spanning frequency at a lag equal to the period. The position of this ridge is the cue to pitch.

1.3. Applying the model to F0 extraction
The aim of this paper is to verify experimentally whether splitting a speech signal over a filter bank offers any advantage for speech F0 extraction. It is important to stress that we do not aim to reproduce all aspects of the perception model in the extraction method. The perceptual quality of called pitch is not the same as for speech fundamental frequency (often called pitch) and the tasks of extracting the former or the latter are not equivalent.

2. METHODS

2.1. Database
Data was taken from an F0 database developed at ATR [19, 20]. The speech sample was at 12 kHz with 16 bit resolution, and labeled for pitch by a crude cepstrum method followed by manual correction. The database contains 500 sentences, read by one male speaker, of which 20 "difficult" sentences were selected and carefully re-labeled by hand. The sentences comprise approximately 19000 voice frames at a 400 Hz frame rate. The F0 values cover a 2-octave range centered on about 125 Hz.

2.2. AMDF
All improvements are based on the Average Magnitude Difference Function (AMDF) method [21]. The AMDF is defined as:

\[ \text{AMDF}(\text{lag}) = \int_{\text{window}} |S(t) - S(t + \text{lag})| \, dt \]

The lag at the first major dip indicates the period. The AMDF produces as a byproduct a parameter that can be interpreted as a measure of periodicity. This is defined as:

\[ \text{PM} = \log_2 \left( \frac{\text{mean(AMDF)}}{\text{AMDF(period)}} \right) \]

The periodicity can be used as a measure of "confidence" in the period value produced by the AMDF algorithm, and also to select channels of high periodicity.

2.3. Evaluation
The AMDF search was constrained to search within 30% of the period specified in the database. The lag at this minimum, the periodicity measure, and an error code are output for each frame. The error code indicates whether the algorithm would have been successful without constraint. It distinguishes subharmonic errors which are not counted as errors in this paper. A "baseline" record of these parameters was derived for the database using standard AMDF. Evaluation was done by frame-to-frame comparison to this baseline. Care was taken to preserve the alignment of processed data: signal smoothing was performed with symmetrical windows, and the outputs of the revcor filters (see below) were shifted in time and phase-adjusted so that the peaks of the envelope and fine time structure of their impulse response coincided with the time origin.

2.4. Revcor filter bank
The experiments use a filter bank program [22] that approximates peripheral auditory filters as "revcor" (or "gammatone") filters, defined by their impulse response:

\[ h(t) = \exp(-T_f t) / T_f \left( 2\pi T_f (0.75 - T_f) \right) \]

where F is the characteristic frequency, \( T_f \) is a latency, \( T_f \) is a time constant of decay, and v is a factor that governs the "symmetry" of the impulse response. The bandwidth parameter was derived from psychoacoustical masking data [23]. Physiological data indicate bandwidths up to three times larger [24, 25]; this factor is explored in the experiments. Bandwidths were set at 1 (standard), 2, 4 and 8 ERB (Equivalent Rectangular Bandwidths) [23]. The filter produces 25 channels uniformly spaced at 1 ERB intervals from 40 Hz to 4000 Hz.

fig. 1. Error rate as a function of center frequency for various channel bandwidths measured in ERBs.

218 219
3. EXPERIMENTS

3.1. Baseline

The error rate of "vanilla" AMDF over the database is 3.84%.

3.2. Individual recorv channels.

The error rates are displayed in Fig. 1 for several bandwidth settings. The rates at 1 ERB bandwidth are very high (around 50%), for other bandwidths they are more reasonable. Rates are lower than baseline in low-frequency channels, and higher in high-frequency channels. The rates at 8 ERB are not very different from baseline, a result which was to be expected given the rather wide filters.

3.3. Half-wave rectification and low-pass filtering.

A possible cause for less good rates in high frequency channels is that it is harder to "register" the fine waveform structure of successive periods. In the auditory system much of this detail is lost, because of the fall-off of synchrony from 1 to 5 kHz [26], an effect similar to smoothing. To check the possible benefit of this effect, the reverber output channels were half-wave rectified and smoothed by convolution with a 20 ms rectangular window (first zero at 500 Hz). Results show an improvement in high-frequency channels, and a slight degradation in low-frequency channels, perhaps because of the loss of information that accompanies half-wave rectification.

![Fig. 1: Error rates for half-wave rectified reverber filter outputs. Dotted line: rates for raw output.](image)

3.4. Cross-channel integration

There are many ways of combining patterns. Here we report a few:

- **addition of AMDFs**
  - The AMDF patterns for all channels are added before searching for the minimum that indicates the period. Error rate, for 1 ERB bandwidth, is 2.9 %
  - **addition of AMDFs of amplitude normalized channels**
    - The reverber filter channels are amplitude normalized (by division by the mean magnitude over a centered window) to give each channel the same weight. Error rate for 1 ERB bandwidth is 5.15 %.
    - **addition of AMDFs of half-wave rectified, smoothed channels**
      - Error rate for 1 ERB bandwidth is 2.7 %.

4. DISCUSSION

At a bandwidth of 1 ERB the error rates are high, probably because resolution of partials prevents interpretation at the fundamental. Rates are much lower at wider bandwidths, particularly for low-frequency channels, which suggests that periodicity information is somehow "better" in these channels. This interpretation is confirmed by results for low-pass filtered speech (table 1).

<table>
<thead>
<tr>
<th>Table 1: Error rates for various degrees of smoothing</th>
</tr>
</thead>
<tbody>
<tr>
<td>window size</td>
</tr>
<tr>
<td>zero at</td>
</tr>
<tr>
<td>error rate (%)</td>
</tr>
</tbody>
</table>

Given this simple result, one might be tempted to apply low-pass filtering systematically. This would be unwisely for a number of reasons. For one, the optimum cutoff frequency depends on the pitch range, and a good setting in one case might be disastrous in others. For another, some applications call for pitch extraction of high-pass filtered speech (such as telephone speech), in which case there is evidently no benefit in low-pass filtering. A more robust answer to be combined information across channels. Simple addition of AMDF patterns yield 2.9 % errors for a 1 ERB bandwidth. This is in striking contrast with the results obtained in individual channels (Fig. 1). Better still is the rate for summed AMDF patterns of half-wave rectified, smoothed channels (2.7 % for 1 ERB bandwidth). Unify weights for all channels, as obtained by amplitude normalization, proved disappointing (5.15 % for 1 ERB bandwidth).

CONCLUSION

An 10 extraction method based that splits the speech signal over a filter bank before calculating the AMDF within each channel and combining the patterns improves reliability of the AMDF method. Future work will examine more sophisticated schemes, such as weighting each channel according to its periodicity measure. More complex algorithms can also be used, such as the channel selection algorithms used by some multi-source separation models [27, 28].

ACKNOWLEDGMENTS

Part of this work was carried out at ATR Interpreting Telephony Research Laboratories, under a fellowship awarded by the European Communities STP programme in Japan. The authors wish to thank ATR for its hospitality, and the CNRS for leave of absence. Special thanks is due to John Holdsworth and Roy Patterson who made available the reverber filter software.

BIBLIOGRAPHY

A COMPUTER ASSISTED METHOD OF INVESTIGATING INTONATIONAL CORRELATIONS IN ADJACENT UTTERANCES

SAPPOK, Ch., KASATKINA, R., RODZASOV, S.

RUHR UNIVERSITÄT BOCHUM, BRD
INSTITUT RUSSKOGO JAZYKA AN SSSR, MOSKVA

0. Abstract
The intonation of adjacent turns in dialogues conveys information of at least two types: it indicates the sentence type of the utterance and in addition the individual attitude of the speaker towards the propositions or parts of them. Is this information subject to direct mapping between prosodic and modal categories? Or is it the result of a process of complex inference? Experiments show that the choice between these alternatives or their combination depends on the communicative task.

1. The Problem
The multiple functions of intonation in a linguistic model can be classified into two subsets: The one captures the assignment of sentence type (question, assertion, exclamation etc.), the other signals the various attitudes of the speaker towards the propositional content of the utterance - something which results in a vast, open class of illocutionary forces.

More research has been done in the second sector than in the first; the functions of which are fewer in number; they are conveyed additionally by means other than intonation. Thus it seems impossible to formulate tasks for empirical investigation. The second field, which we will call the subjective modality, seems to be subject to individual variation; the number of categories is unknown. Indeed, it seems questionable whether they are categories at all.

In this paper we present a method of experimental research in this second area, making use of qualitative information to make an entire communicative situation repeatable and subject to quantification, in a way similar to the procedures formulated by OLIVER and GARDENBURG 1983. The evidence we will adduce will prove to favour one of the following alternatives: Can we assign the modal categories postulated directly to an utterance and its intonation contour? Or is the interpretation the result of a complex process of inference? Furthermore, what kind of information seems necessary? A similar alternative has been formulated by LEVINSON 1983 under the heading conversational analysis.

The first result of our experiments, making use of a non-quantitative interpretation, shows evidence for the inferential model. As to the set of information to be used, there seems to exist a high degree of variation; even in the absence of sufficient information, the modal utterances and their adjacent combinations are interpretable, since there appears to be a set of "default" knowledge.

2. The Method
The material consists of 12 microdialogues consisting of 4 turns each, and a preceding description of the situation. As to the organization of the material in the form of a data base cf. SAPPOK 1990. The situation consists of a variable combination of propositions. The dialogues having always the same lexical form, as can be seen in the samples shown in chart 1.

The description of the situation and the text of the dialogues were presented visually in written form to pairs of Russian native speakers who performed them according to the instructions. The resulting utterances were digitalized and reorganized for the user in the form shown in Chart 2, making use of the computer program developed by KNIPSIL'D 1990. The display shows the instruction categories in symbolic form; Ivanova's prior behaviour has been good (poly vymyty) or bad (kleenka isporchena), the assignment of turns to the speakers changes from S to R and vice versa. The following symbols show keys to be pressed, after which the resulting dialogues can be heard.

Стурка 1.1.
А. И В. хорошо отвечает Иванова.
А. хочет усилить его отношение.
А. - Ты замечал, что полы вымыты?
В. - Да, а что это сделано?
А. - Иванова.

Стурка 1.2.
А. хорошо отвечает к Ивановой, а В. плохо. А хочет выразить отношение В. к Ивановой в хорошем.

А. - Ты замечал, что полы вымыты?
В. - Да, а что это сделано?
А. - Иванова.

А. - Иванова.
А. - Иванова.
А. - Иванова.

А. - Иванова.
А. - Иванова.
А. - Иванова.

А. - Иванова.
3. The Experiments
The instruction is assumed to determine the intonation of the turns. Various combinations of the turns and descriptions of the situation are used to construct of stimuli to be presented to the subjects. We shall describe in detail two experiments representing two extreme positions, i.e. maximal and minimal information on the basis of which the subjects have to make their decisions.

In the first type of experiments, the combination of the situation description and the dialog is presented with the exception of one detail - the presumed opinion on Ivanova as bad or good. It is this 'opinion' or 'attitude', which is to be extracted on the basis of the intonation of A. or, in a separate experiment, of B. A similar task is the reconstruction of Ivanova's pre-dialog behaviour of Ivanova.

The second type of experiment utilizing isolated utterances (turns) presents the subject with the task of determining the similarity of intonational contours of the repeated answers, the type of question between them (weak or strong), and the degree of emotional expression.

In the third type of experiments the subject has to take part in the dialog himself, uttering responses to the computer in turn. The subject is given the possibility of hearing the dialog and of repeating it as often as necessary until he finds it adequate, making use only of the information conveyed by the intonation which he is reacting to. Chart 3 shows

Chart 3. a) - c) Three reactions of a subject to neutral, positive and negative utterances in the dialogue with the computer.

3.1. Type 1: two questions of type A 2 as reactions to B 1 utterances of neutral, positive and negative versions.

3.2. Type 2: three questions of type A 1 as reactions to B 1 utterances of neutral, positive and negative versions.

4. The Interpretation

In determining the speaker's attitude subjects show in some cases a high degree of similarity, while in other cases their interpretation remains disparate.

The overall picture is the following:

- Neutral attitude is recovered with greater accuracy in the context of positive behaviour; it seems difficult for the speaker to remain neutral in the context of negative behaviour.
- In the case that behaviour and attitude have different values, subjects have difficulty recovering the original intentions.

The intonation seems to convey not the isolated speaker-generated values, but rather the combination of a negative attitude and negative behaviour usually results in positive answers! This can be an expression of satisfaction resulting from the perception that the judgments correspond.

These results show that there is no set of modal features that can be interpreted in isolation. The modal cues in the intonational contours must therefore be interpreted in combination with various other types of information.

Additional evidence in favour of this kind of model is to be found in the results of experiments of Type 2.

Comparing the repeated answers B 1 and B 2, (made comparable by cutting off the initial "da" of the latter) subjects reveal the highest degree of dissimilarity in dialog 2.2, where speaker B. tries to influence speaker A., knowing that the latter's attitude towards Ivanova is contrary to his own. The intervention question A.1. seems to be a signal to speaker B. that his attempt to influence A. was not successful and has to be repeated with a modified intonation.

The judgment "neutral" always implies a high degree of similarity in the case of 2.2., a less intense degree, in the case of 2.3. and 2.1. corresponding to a decreasing need for resistance.

The exact mechanisms of modal expression and interpretation must remain open until the results of quantitative, statistical analysis are available.

Preliminary interpretation shows that:

- the reaction of the subjects to the situations and dialogues is not random;
- the interpretation is the result of a process of inference, taking into account different types of information;

- even in the case of the absence of exact information an interpretation still seems possible; in this case a "default" standard situation seems to be assumed.

References:
Early Germanic was a 

prosodic unit that goes back to so-called syllable accents have been attested only in North Germanic: in Swedish, Norwegian, Danish, and in the Rhein-Limburg area. If we agree to view the glottal stop and preaspiration as analogues of stød, our map will include English, Icelandic, Faroese, and several additional dialects of Dutch and German, but its borders will not move more to the south. All other accents can be reconstructed only from the traces they left on vowels and consonants. However, if the place of ancient stress is partly deducible from the reflexes of diphthongs and triphthongs on the vast territory from Friesland to Lustenau, the type of old stress and the number of the once relevant accents remain a matter of speculation. Combining the data supplied by Verner’s Law and Akzentungsgruppen (a process responsible for the variation of the /a/ type), we can state that stress in Early Germanic remained movable within a bimoric complex long after it became fixed on the root. Some accent-like units most probably existed in North Germanic about two millennia ago, but it does not follow that they were present in the languages of the Germanic tribes south of Cologne.

To the extent one can judge by the situation in the Rhein-Limburg area, accents delimited certain types of bimoric bases and performed the function of boundary signals. The prosodemes of the Swedish-Norwegian type (accents 1 and 2), governed as they are at present solely by the number of syllables rule, could not be the prototypes of such accents. Accents 1 and 2 (with the exception of a few dialectal occurrences) do not depend on the phonetic basis, and therefore it is reasonable to assume that this independence is late. In Danish, stød and /a/ are connected with the basis and with the (actual) number of syllables in a word. In German and Dutch dialects, the appearance of correption and its opposite is also subject to the prosodic basis and the (original) number of syllables: apocopated words are accent differently from nonapocopated. In both areas, the basis is the older distributional factor, the only one that existed prior to apocope. Danish and Germanists regard stød as a late prototype.

One of the implications of their theory is that Danish stød and WG correption are unrelated, which also follows from the evidence on the chronology of stød untatable. The WG analogue of stød distinguishes between open and close vowels. According to the nasalized Rigaan pattern, correption occurs on the reflexes of the open vowels /a:/ e: / and of the old diphthongs, insofar as they were smoothed. Words of this group are said to have spontaneous correption. The reflexes of old /i:/, /u:/ and monomorphized diphthongs are corrected when the word is dysyllabic or apocopated and when the postvocalic consonant is voiceless. In dysyllabic and apocopated words whose root consists of a short vowel followed by a resonant and an obstruent, i.e., in words with diphthongal groups, correption is also possible only before a voiceless obstruent, so in Hunde but not in Kants.

The vowels /i:/ u:/ do not belong with /a:/ e: / because in WG they were treated as diphthongal groups, namely, as /i:/ u:/, and so forth. Correption marked the end of the bimoric monosyllabic basis. All the early Germanic languages were more or less, and, as evidenced by Akzentungsgruppen, could fall on either mora or a bimoric complex; correption separated the root of the word that served as the locus of shifting stress. In words with diphthongal groups, correption occurred only before a voiceless obstruent because a voiceless obstruent marked the end of the prosodically active string by its voicelessness. Diphthongs were accent like diphthongal groups: when smoothed, they did not differ from the other long open vowels, and when preserved as units with two distinct elements, their role /i:/ was donated.

In our classification of phonemes, we often try to discover whether Early Germanic obstruents were phonologically voiced/voiceless and diphthongal. As it may well be that a distinctive feature is a more complex phenomenon than we think. If we treat distinctive features pragmatically (“what do they do in the system?”), rather than as mere classificatory labels, /p t k/, to give one example, can be strong from the view of syllable contact and voiceless in being able to delimit a certain type of basis. Later one of the functions can disappear and then voicelessness might remain the only feature of /p t k/. Still later even this feature can become detrimental to the performance of the consonants’ next role, and then aspiration (reinforced by the new circumstances) will assert itself, and so forth.

Diphthongal groups (including /i:/ u:/), as well as old monosyllables with a combinatory
basis, had no correction before voiceless consonants, and it is not known how these words were pronounced. Two situations can be imagined. In some cases, noncorrupted words probably had "nothing." The opposite of Danish stød, no-stød, is the negation of stød, and foreigners do not regard it as a special prosodeme. The intuitive impression is that stød is marked and no-stød "unmarked" and that the opposition is privative. But it is also probable that the opposite of correction was itself an independent boundary signal within the framework of an equipollent opposition. If correction presupposed increased energy of articulation and shortening of the vowel, its opposite could have been associated with the general relaxation of the vocal tract and lengthening of the phonetic basis. It, too, could have been realized as a short break, but smooth and breathed, rather than abrupt, when the vocal chords are constricted or compressed. Given two full-blown boundary signals, we can perhaps explain the origin of Scandinavian prespiration. The distribution of prespiration in Icelandic and Faroese is almost the same as that of the glottal stop in Cockney and the West Jutland stød. It is tempting to suggest that prespiration is related to stød as sleeplessness is to snore and that at one time prespiration was the "mirror" opposite of stød. 

A difficult problem confronting us in areas in which correction and "extension" are distributed according to the "mirror rule," as compared to the Riparian one: words with the reflexes of ja: e: o/ and of smoothed diphthongs do not have correction, and in the other cases, correction occurs before voiceless, not before voiced, obstruents. In most of these vernaculars, correction is phonetically weak, whereas the extending accent is prominent. The riddle of the "mirror rule" will remain insoluble if we keep looking on correction as the only thinkable marker of old bimoric bases. If, however, we accept the possibility of gradual change in the system - [MM] (two morae and correction) or [MM'] (two morae and a pause) - the Riparian and the rule of the peripheral dialects from northern Limburg to Arzbach will emerge as equally probable. The unmarked signal has a blurred realization everywhere; in Riparian, the opposite of correction is "nothing," in Klevo, Arzbach, etc., the opposite of "extension" is a weak shadow of forceful correction.

It cannot be stated whether the two ancient boundary signals always or at least sometimes formed an equipollent opposition. In Danish, no-stød is never marked; yet as a theoretical possibility it may indeed arise, since we might define the recipient of the signal: the opposite of a boundary signal is "nothing," in Klevo, Arzbach, etc., the opposite of "extension" is a weak shadow of a strong accent.

In the Scandinavian languages, stød (correction) never marks apocope, but in West Frankonian it regularly does thus being a feature of the old signal. In old monosyllables with spontaneous bases, correction, indeed, has nothing to do with the occurrence of the new signal. In apocopated words it is an analogue of the two-peeked accent. Apocope endowed on a boundary signal with a new role, and its yield increased. Our ideas of phonological relevancy are still crude. When in certain dialects stød occurs only in monosyllables, and no-stød only in disyllables or when stød is allowed before voiced consonants and no-stød before voiceless ones, we conclude that the units under consideration are redundant or that they belong to usage rather than the system. Complementary distribution is interpreted as redundancy. This is an unacceptable approach in phonetics (3) and even more evidently so in prosody. The two boundary signals would not have merged if they had had no use, and the role of apocope enhanced the unit's visibility. From an acoustic point of view only "extension" resembles the circumference of northern Saxon dialects, whereas any signal of apocope comes close to or merges with the circumference of general phonetics, and it is no wonder that both "extension" and correction are often perceived as two-peeked: the boundary signal that became the marker of apocope changed its realization under the influence of the new function. Even if the original opposition ('1) - ['1] was equipollent, the loss of endings turned it into a privative: the boundary signal was chosen as the accent of apocope and became the opposition's marked member and the most easily discernible periodic shibboleth of the entire prosodic system. Frings carried his point too far when he insisted on the equal importance of correction and "extension" in Low Franconian, but even less convincing is the thesis of Dutch dialectologists that "extension" is marked in Limburg because Dutch pronunciation is in general smoother than German. Markedness is a functional concept and cannot be derived from the articulatory base.

The Danish spontaneous and combinatorial accentuation are seldom distinguished. Only in East Jutland does one come across e.g. with stød and e. no-stød without stød (i.e. in monosyllables before a voiced and a voiceless obstruent respectively). It is more probable that Danish dialects simplified ancient diversity than that we developed the juxtaposition of two spheres, but there could always have existed more and less complex systems.

It seems that in the epoch following the fixing of stresses on the root the Germanic languages of the North made one of two boundary signals (abrupt and smooth) depend on the type of phonematic basis. These signals acquire greater importance when they can be associated with apocope and when the number of syllables rules across. We extend evidence to the existence of such a distinction (stød) and extension in all the Early Germanic dialects, and there is no evidence to the accents registered in Old Indian, Ancient Greek, and Balto-Slavic. Especially unproductive is the discussion about dynamic stress versus musical stress, for these concepts have foundation in either phonetics or phonology.
TIMING IN CATALAN

D. Recasens

Institut d'Estudis Catalans, Barcelona, Spain

ABSTRACT

This study is a preliminary analysis of timing organization in Catalan. The topics under investigation are compression of stress groups and stressed syllables, final lengthening, and rhythmic alternation in unstressed syllables.

1. INTRODUCTION

Recent phonetic studies on speech timing disclaim a strong version of the opposition between syllable-timed (i.e., Spanish, French) and stress-timed (i.e., English, Swedish) languages. There is little evidence (if any) for isochrony within the syllable or foot domain in the two language types; instead syllable and foot duration appears to increase as a function of segmental complexity.

To cope with this variation, finding two alternative views have been proposed. Some scholars [2, 10] believe that languages are perceived as syllable- or stress-timed because of phonological factors. Thus, in contrast to syllable-timed languages, stress-timed languages allow complex consonant clusters in word position, and may reduce all vowels to schwa in unstressed word position. Moreover, the addition or suppression of schwa affects syllabification in the former (i.e., French) vs the latter language group.

Phonologists are however reluctant to abandon acoustical or metrical timing measures. It seems now well established that there is no clean dichotomy between the two language types. Moreover, rhythmic differences among languages probably reflect the contribution of several durational and spectral constraints [8].

In this paper I will look for phonetic correlates of timing organization in Catalan. In spite of Catalan being a Romance language, its phonological makeup does not fully agree with that of other syllable-timed languages such as Italian or Spanish. Indeed Catalan allows consonant clusters up to three segments in syllable-final position and has a schwa in unstressed position. Differently from English, Catalan [a] always behaves as a syllabic nucleus (as in French). Because of its particular phonological structure, Catalan is a good candidate to test the interaction of phonetic and phonological factors in the rhythmic structure of languages.

2. METHOD

Three Catalan speakers were asked to read a list of nine nonsense words. In order to preserve naturalness in the reading task each nonsense sequence was uttered after a meaningful Catalan sentence with the same stress pattern and syllable structure. The nonsense words were preceded by the stressed monosyllabic Catalan word "a" ("he says"). They were composed of one stressed syllable (pa) and zero, one or two preceding and/or following unstressed syllables (pa) (see Table 1). Schwa can appear in unstressed position in Catalan.

Several segmental units were measured from waveform displays, namely, stress group (a stressed syllable preceded or followed by 0, 1, or 2 unstressed syllables), vowel (stressed [a] and unstressed [a]), and consonant (stressed and unstressed [p]).

3. RESULTS

3.1. Stress group durations

Measurements show a monotonical increase in stress group duration with the number of syllables within the group for all sequences. The two variables are highly correlated (r = 0.9, I and 1 according to speaker). This is exemplified in Figure 1 which displays durations of one-, two-, and three-syllable size stress group intervals according to speaker Re.

A linear correlation between the two variables is not exclusive of syllable-timed languages (French: 5; Italian: 10) but has been documented in stress-timed languages as well (Dutch: 10; English: 9).

3.2. Final lengthening

There is very scant evidence in support of the hypothesis that syllable-timed timing is incompatible with final lengthening. Final lengthening has been reported to occur in French [5], Spanish and Japanese [7]. It does not show up in vowels in Italian stressed syllables and vowels [12].

Final lengthening in Catalan was calculated separately for stressed and unstressed syllables, vowels and consonants. In all cases it was equated to the ratio between average durations in final vs medial position.

All speakers show robust final lengthening effects for unstressed syllables (i.e., stressed vs unstressed vowels, vowels and consonants [English: 9; Italian: 12], and for stressed and unstressed vowels vs consonants [French: 5]).

Stressed and syllable-timed languages may differ in the magnitude of the lengthening effect. In support of this hypothesis there is less stressed vowel final lengthening in Catalan (38%, 22% and 24% vs (all other speakers) than in English (50%) [6] oxytone vs paroxytone sequences.

3.3. Compression of stressed vowels and consonants

In comparison to syllable-timed languages, stress-timed languages are expected to show a higher degree of compression of stressed syllables duration as a function of the number of unstressed syllables within the stress group. Moreover sensitivity to compression effects may depend on whether the unstressed syllables precede (carryover compression) or follow (anticipatory compression) the stressed syllable.

Significant anticipatory effects at the p<.01 level were found for stressed (a) when the number of following unstressed syllables increases from 0 to 2 in all sequences (i.e., [pa] vs [papa], [papa] vs [papap], [papap] vs [papapapap]) for two speakers and in only one of those three sequence types for the other speaker.

Consistently with data from the literature, there is less anticipatory compression for consonants than for vowels since it only occurs in the 1 vs 2 syllables condition when no syllable precedes the stressed syllable (i.e., [p] vs [pap]) (all speakers).

Carryover effects on vowel and consonant duration are only significant in some cases when the number of preceding syllables increases from 0 to 1 and no syllable follows the stressed syllable (i.e., [pa] vs [papa]).

Figure 2 illustrates anticipatory and carryover compression effects for stressed syllables according to speaker Re. The figure shows much less stressed syllable shortening than much less stressed vowel shortening) in the 2 vs 1 than in the 0 vs 1 following syllables condition. In particular stressed vowels in polyphones are shorter than those in paroxytones by 13%, 11.5% and 10% according to speaker.

Data for Catalan presented here are somewhat consistent with those for other stress-timed languages showing larger and more robust carryover compression effects and thus suggesting the existence of a left-dominant foot structure (Swedish: 6; English: 6). Concerning syllable-timed languages a similar trend has been found for Italian [13]. Other stress-timed languages show no anticipatory compression effects (Japanese: Spanish: 7), or do not favor right-to-left compression trends (Spanish: 11).

3.4. Unstressed syllables

Statistical analysis on unstressed syllables duration allows drawing the following conclusions:

(a) differences in duration among unstressed syllables are not larger than 8% to 10% of the mean unstressed syllables duration;

(b) for all speakers pretonic unstressed syllables are located two syllables away from the stressed syllable (i.e., word-internal unstressed syllables) are the shortest of...
all unstressed syllables in the word;
(c) for two speakers posttonic unstressed syllables which are adjacent to the stressed syllable are particularly long. The fact that durational differences across unstressed syllables are particularly small conform better to a syllable-timed than to a stress-timed language model [see 3 for discussion]. Moreover, Catalan unstressed syllables show a rhythmic pattern which has also been reported for other syllable-timed languages, with weak initial unstressed syllables and strong medial unstressed syllables (more so if immediately posttonic). Indeed unstressed syllable duration in Spanish and Japanese decreases in the progression final—medial—initial [7]; moreover it has also been found for French that two pretonic unstressed syllables should conform to a weak-strong (W-S) pattern [3]. Stress-timed languages usually show significant shortening of unstressed syllables next to a stressed syllable [7]. Therefore in languages of this group syllable duration within the word decreases in the progression initial—final—medial [Swedish: 1; English: 7]. Italian researchers have also found shorter unstressed syllables in word medial vs absolute initial position in Italian [4].

4. SUMMARY
Analogously to syllable-timed and stress-timed languages Catalan shows final lengthening and a stress group duration which is proportional to the number of syllables within the group. Differently from syllable-timed languages such as Spanish, Catalan appears to favour anticipatory vs carryover compression of stressed vowels within the stress group; analogously to Italian this trend is probably weaker than in stress-timed languages. Like other syllable-timed languages, positional realizations of [a] differ little in duration and shorten when adjacent to unstressed syllables but not to stressed syllables.

ACKNOWLEDGMENTS
This work has been supported by the ESPRIT-ACCOR project (BRA Action 3279) from the European Community.

5. REFERENCES

TABLE I. List of nonsense words used in the experiment.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[p a]</td>
</tr>
<tr>
<td>2</td>
<td>[p a p a]</td>
</tr>
<tr>
<td>3</td>
<td>[p a p a p a]</td>
</tr>
<tr>
<td>4</td>
<td>[p a p a]</td>
</tr>
<tr>
<td>5</td>
<td>[p a p a p a]</td>
</tr>
<tr>
<td>6</td>
<td>[p a p a p a]</td>
</tr>
<tr>
<td>7</td>
<td>[p a p a]</td>
</tr>
<tr>
<td>8</td>
<td>[p a p a]</td>
</tr>
<tr>
<td>9</td>
<td>[p a p a p a]</td>
</tr>
</tbody>
</table>

FIGURE 1. Stress group duration as a function of the number of syllables (speaker Re). The data are represented separately for oxytone (continuous line), paroxytone (dashed line) and proparoxytone (dotted line) nonsense words.

FIGURE 2. Anticipatory (upper graph) and carryover (lower graph) compression of stressed syllables as a function of the number of unstressed syllables in the stress group (speaker Re). The data are represented separately for one (continuous line), two (dashed line) and three (dotted line) preceding (anticipatory compression condition) and following (carryover compression condition) syllables. Significant compression effects are marked with an asterisk.
THE TIMING OF VOWEL AND CONSONANT GESTURES IN ITALIAN AND JAPANESE

Caroline L. Smith
Yale University and Haskins Laboratories, New Haven, Connecticut

ABSTRACT

Languages commonly described as syllable-timed, such as Italian, are perceived as having a rhythm in which each vowel is the nucleus of a rhythmic unit. In contrast, in mora-timed languages such as Japanese the basic rhythmic unit, the mora, depends on the durations of both vowels and consonants. It is proposed here that the basis for the contrast between these two types of languages is a correlation between the temporal organization of articulatory gestures for vowels and consonants and the role of vowels in the overall rhythm of a language: in syllable-timed languages, vowels have primacy over consonants, but in mora-timed languages vowels and consonants are of equal importance.

1. INTRODUCTION

The hypothesis being tested is that stress, syllable- and mora-timed languages are characterized by more or less independence in the timing of vowel and consonant gestures. The term gesture refers to an abstract, dynamic unit associated with the production of a particular vowel or consonant that controls the spatiotemporal movement of one or more articulators towards a target. Both of the models of gestural timing that will be compared here assume that the temporally overlapping production of gestures is responsible for their apparent context dependence, but the models differ in their accounts of how gestures are coordinated in time. Both models were proposed to account for English and other languages, but seem to capture characteristics of different types of rhythmic behavior.

The vowel-based timing model [5, 6, 8] claims that gestures for vowels and consonants are coordinated at different levels. Vowel gestures are coordinated with respect to one another, and consonants are coordinated with respect to the vowels. This model predicts that vowel gestures will be unaffected by temporal changes to consonant gestures. Because of the primacy of vowels in determining syllable-timed rhythm, the vowel-based model was expected to apply to Italian.

The vowel-and-consonant timing model [2, 3, 4] claims that vowels and consonants are coordinated at the same level. Since this means that intergestural timing for vowels and for consonants is independent, a timing change to any gesture is predicted to cause adjustments in both sets of gestures. This model was expected to apply to Japanese, because mora-timing requires the temporal integration of vowels and consonants. In Japanese, the timing of two vowels relative to each other is not as predictable as the sequencing of individual vowels, which is determined by changes in the duration of intervocalic consonants.

Because the two models' predictions differ primarily in the extent to which vowels are affected by changes in the timing of consonants, contrasting utterances that differ only in the length of an intervocalic consonant provide a way to compare the two models. However, since the predictions of the models are couched in terms of abstract gestures, the gestures, in order to be compared experimentally, must be associated with specific articulatory movements. Vowel gestures, for example, can be associated with an appropriate movement of the tongue body (or root), and consonants with the lips or the tongue forming a constriction in the supralaryngeal part of the vocal tract. Associating gestures with movements in this way makes it possible to compare gestures in different contexts, but it does not differentiate the roles of the various articulators in making the construction.

2. METHOD

In order to measure the movements of the tongue associated with vowel gestures, as well as the lips and jaw, data were collected at the NIH X-ray microbeam facility at the University of Wisconsin. The microbeam records the movements of the tongue, lips and jaw by means of microscopic X-rays tracking tiny gold pellets attached to the articulators [1]. Pellets were attached to the lower lip and upper incisor (to correct for head movement), lower incisor (to measure jaw movement), lower and upper lip, and to four points along the midline of the tongue, starting approximately 1 cm behind the tip of the extended tongue. The microbeam data consist of the horizontal and vertical trajectories of each of these pellets.

The present paper is a subset of a larger study, in which three native speakers each of Italian and Japanese participated. Data from only one speaker of each language will be discussed in this paper. Each speaker produced, in carrier phrases designed to be comparable across languages, syllabic utterances of the form "m+CV2" where V1 and V2 were /i/ or /u/, and C was one of /b/, /p/, /w/, /l/, /m/, /n/, /m/ or /n/. In Japanese, utterances with /u/ or /u/ as the intervocalic consonant were also included, as were some sets of gestures that were excluded because /l/ palatalizes in this context. The Italian speaker produced 9 to 11 tokens of each utterance, and the Japanese speaker produced 16 to 17 tokens. The movement trajectories were digitized and smoothed prior to analysis.

Because of the very high correlations among the two tongue pellets (as high as .95 between the X-dimensions of two pellets), a factor analysis was performed on the X and Y positions of the pellets at successive 50 ms frames with the intention of extracting factors that would reflect the positioning of the tongue for the various vowels. Examination of the movement trajectories had suggested that the frontmost tongue pellet showed primarily movement associated with the alveolar consonants, so it was excluded from the factor analysis, leaving 6 dimensions, from which 2 factors were extracted. The first of these was primarily associated with horizontal movement, and the second with vertical movement. Pellet trajectories were also measured that were expected to show movement typical of specific gestures. The trajectories that were measured are shown below.

<table>
<thead>
<tr>
<th>Trajectory</th>
<th>Associated Gesture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower lip vertical</td>
<td>/l/ as intervocalic consonant</td>
</tr>
<tr>
<td>Tongue Tip vertical</td>
<td>/a/ as intervocalic consonant</td>
</tr>
<tr>
<td>Tongue Dorsum horizontal</td>
<td>/u/ as intervocalic consonant</td>
</tr>
<tr>
<td>Tongue Body Rear vertical</td>
<td>/u/ as intervocalic consonant</td>
</tr>
<tr>
<td>Horizontal tongue factor</td>
<td>/w/ as intervocalic consonant</td>
</tr>
<tr>
<td>Vertical tongue factor</td>
<td>/u/ as intervocalic consonant</td>
</tr>
</tbody>
</table>

The utterances measured were those in which the two vowels were different, as these permitted the identification of movements from the first vowel to the second. Five time points, defined as the edges of periods of zero velocity, were located in each of the trajectories associated with vowel gestures: the onset of movement towards the first vowel, the time at which the movement for the first vowel reached its target, the end of the plateau region for the first vowel, the time at which the movement for the second vowel reached its target, and the end of the plateau region for the second vowel.

3. RESULTS

Different intervals between the labelled time points were measured in order to determine whether the time between the two vowels was changing when the length of the intervocalic consonant changed. ANOVAs were run for each subject separately, with the intervals between the labelled points as dependent variables and grouping
Factors Length (of intervocalic consonant), Place of articulation, Consonant Identity, and Vowel quality.

Figures 1 and 2 illustrate tokens of /mipa/ (solid line) and /mippa/ (dotted line) from Italian and Japanese, aligned at the release of the initial /p/. In Italian (Figure 1), the large bumps associated with the production of /a/ in the top three articulatory trajectories are virtually identical in /mipa/ and /mippa/, the rear and downward movements for /a/ also coincide. The two utterances differ in the positioning of the central hump in the Lower Lip trajectory, which corresponds to the intervocalic consonant, relative to the other movements. The raising of the lower lip for /p/ occurs earlier relative to the preceding lip movement (p<.001 for the effect of Length) and to the tongue movement for the /a/ than does the raising for /p/ (p<.001), resulting in the preceding vowel being shorter acoustically before the geminate (p<.001), a well-known characteristic of Italian [7].

Figure 2, for Japanese, shows the raising of the lower lip for the intervocalic consonant occurring at about the same time relative to the preceding lip and tongue movements, with the preceding vowel slightly longer acoustically preceding the geminate (p<.001). Although the tongue raising and fronting begins at approximately the same time in both utterances, the lowering and backing for /a/ is significantly delayed when following /p/ (p<.001 for the effect of Length).

This impressionistic pattern is borne out by measurements of the interval between the times at which the two vowels reach their targets, whose approximate locations are indicated by arrows on the figures. This interval was consistently longer in Japanese (p<.001 for all trajectories). The statistical results for Italian were more variable, but with negligible numerical differences found in the contexts of the two consonant lengths. These results support the hypothesis that the second vowel is delayed relative to the first in Japanese but not in Italian.

Although preliminary, the results shown here do suggest that the timing of the two vowels relative to each other is controlled independently of the consonants in Italian but in conjunction with them in Japanese. The most immediate implication of this is that either model of timing organization can claim to be the most insightful for both types of languages. The apparent relation between the form of temporal coordination between vowel and consonant gestures and the corresponding differences in linguistic rhythm suggests that the organization of gestural timing may be a source for the differing rhythmic behavior between syllable- and mora-timed languages.

4. REFERENCES

Work supported by NSF grant BNS-8828099 and NIH grant DC-00121 to Haskins Laboratories.

The factor analysis was a principal components analysis using BMDP 4M, with a VARIMAX rotation. The factor scores were then calculated for each frame of data, providing trajectories similar in form to the pellet trajectories.

Results reported here for the effect of Length are based on 1,145 degrees of freedom for Italian and 1,195 for Japanese.

Figures 1 and 2. Productions of, at the top, Italian "Dica mipa molto" (solid lines) and "Dica mippa molto" (dotted lines), and at the bottom, Japanese "Boku wa mipa mo aru" (solid lines) and "Boku wa mippa mo aru" (dotted lines). Time is along the horizontal axis; each tick mark indicates 100 ms. The utterances within each picture were aligned at the release of the initial /p/ in the target word. The times at which the vowels reached their targets are shown by arrows (solid for the single consonant, dotted for the geminate).
PAUSING IN TEXTS READ ALOUD

E. Strangert

Department of Phonetics, University of Umeå, Sweden

ABSTRACT
Perceived pauses in Swedish news texts read aloud were investigated. The pauses were analyzed to determine their distribution as well as their acoustic correlates and the perceptual relevance of these correlates. Most pauses occurred at syntactic boundaries, and the higher the rank of the boundary, the greater the probability of a pause. The acoustic correlates of pauses, in addition to silence, include prepausal lengthening, resetting of intensity and Fo, and voice quality irregularities. In general, the higher the rank of the boundary, the stronger and more varied were the acoustic correlates. Moreover, the data demonstrate that syntax plays a role not only in the production but also in the perception of pauses.

1. INTRODUCTION
This paper reports results from an ongoing project about pausing in Swedish. First, it concerns pauses in texts read aloud. Thus, the analysis only marginally includes hesitations and other phenomena that characterize ordinary speech situations. Secondly, the project combines a prosodic and a syntactic as well as a textual perspective on pauses. The purpose is to describe where pauses occur in relation to language structure, in particular to boundaries of different kinds. The purpose is, moreover, to learn about how these pauses are manifested acoustically, and finally, how the acoustic correlates contribute to the impression of a pause. Thirdly, "pause" in this study means "perceived pause." The focus is on those parts in the speech stream at which a pause is heard. By choosing this rather than an acoustic definition, pauses without a silent interval will not be excluded from analysis. The study includes normal, fast and slow renditions of the texts. A detailed account of the purpose and general outline of the project is given in [13]. Other studies with a similarly wide perspective on pausing include [10, 2, 15, 4].

2. MATERIAL AND ANALYSES
The material consisted of two news cables with a total of 810 words. Some of the original words had been exchanged for specific test words inserted in different syntactic positions to make it possible to study prepausal lengthening at different types of boundaries. The texts were read by ten male speakers. Each one read the material at his normal speed and at a faster as well as a slower speed. All the material was recorded on tape and registered on micrograms. Prior to further analyses, two listeners identified the pauses from the recordings. Of the total number of pauses identified, the interrater reliability varied between 78 and 94% for the different speakers. These percentages may be compared to the 72% reliability in a Dutch study by de Rooij [10]. de Rooij had five persons listening to one speaker which reasonably should give a lower figure.

A syntactic analysis of the texts was also carried out with units such as paragraphs, sentences, clauses and phrases defined as in traditional grammar. The boundaries separating these units were marked as paragraph (§§), sentence ($$), clause (/?) and phrase (?) boundaries, respectively [13].

3. PAUSE DISTRIBUTION
The occurrence of pauses in relation to language structure has been investigated for different languages and conditions. Studies of speech read aloud have been based on German [2], English [15] and Dutch [10, 1].

In the present study some positions seemed to almost obligatorily attract pauses, while in other positions the occurrence of pauses varied for the different speakers. Positions where at least five of the speakers made a pause perceived by both listeners were termed "strong pause positions". All strong pause positions coincide with syntactic boundaries, and as might be expected, all paragraph and sentence boundaries are strong pause positions, independently of speech rate. For clause and phrase boundaries, speech rate is more important. The slower the speech, the more frequent the pauses. Figure 1 shows how strong pause positions are distributed over clause and phrase boundaries.

4. ACOUSTIC CORRELATES
So far, the normal rate data for six of the speakers have been analyzed. Measurements were made of silent intervals, test word durations (to estimate prepausal lengthening), as well as Fo before and after pauses. There was also an evaluation of voice quality irregularities before (prepausal) and after pauses. Figure 2 presents data for silent intervals.

It is apparent that even though the absolute durations vary widely between the speakers, they follow the same pattern. The duration of the silent interval matches the rank of the boundary. If the mean silent interval at paragraph boundaries is given a duration of 1 for each of the speakers, then at sentence boundaries the mean silent interval about .6 and at clause boundaries about .2 of the reference duration. In general the mean silent interval at phrase boundaries is somewhat
5. PERCEPTUAL ASPECTS

The pauses in this study were usually identified when suppon speech data related to the pause positions were collected. This procedure does not permit conclusions as to the perceptual significance of the respective correlates or how they combine into the impression of a pause. (There may also be other relevant correlates than those which were chosen. In fact, it seems that resetting of intensity is such a correlate.)

Figure 2. Mean silent intervals at paragraph, sentence, clause and phrase boundaries. Data for six speakers.

shorter than at clause boundaries, but the differences between these categories are very small [12]. Butcher [2, p 175-179] similarly measured silent intervals between sentences as well as between different types of clauses. As in the present study, the intervals were longer between sentences than between clauses. In addition, Butcher found significant differences of the silent intervals within the clause category.

There is a positive correlation between the acoustic signalling and the rank of the boundary for other pause correlates, too [11, 14]. Fo before a pause tends to drop to a lower value, and Fo after a pause tends to start at a higher value the higher the rank of the boundary. Thus, the resetting is greatest at paragraph boundaries.

Irregularities of voicing, e. g. creaky voice, present a similar pattern. Most pauses with such irregularities occur at paragraph and sentence boundaries, and the higher the rank of the boundary, the stronger the irregularities. However, prepausal lengthening deviates from the general trend. There is no apparent positive correlation between the degree of lengthening and the boundary rank. This fits in with the observation that there is no obvious difference between the lengthening before sentence and a paragraph boundary [8]. Several studies indicate complementarity between lengthening and the following silent interval [4, 5].

6. REFERENCES


RHYTHMIC STRUCTURES IN POETRY READING.

Anita Kruckenborg, Gunnar Fant and Lennart Nord

Department of Speech Communication and Music Acoustics, KTH, Box 700 14, S-100 44 STOCKHOLM, SWEDEN.

Phone 46 8 790 7872, Fax 46 8 790 7854

ABSTRACT

Our study is concerned with the reading of metrically structured verse. We find an apparent tendency of an integration of pauses within and across verse lines to maintain a rhythmic continuity of mean interstress intervals. Similar rhythmic traits have earlier been found in prose reading but with a greater variability of the duration of boundary spanning intervals. Meter specific temporal patterns of strong and weak syllables have been found to comply with expectancies. Thus, a main difference between realizations of iambic and trochaic patterns is that a more precise stress of the iambic weak syllable. This trend in part reflects a difference in syllable complexity, the average number of phonemes in the iambic weak syllable being less than that of the trochaic weak syllable.

1. INTRODUCTION

Rhythm in the reading of a poem may be looked upon both as a literary and as a linguistic phenomenon. The analysis must handle aspects and tools from metrical as well as prosodic points of view. Important are the concepts of meter and rhythm. Today metricians try to distinguish between meter and rhythm in poetry. Meter is the abstract, pure pattern of the alternation of weak and strong syllables. Rhythm in poetry, then, is the product of a delicate interplay between this abstract pattern of meter and the normal rhythm of language in prose. Thus, in the reading of a poem, meter can only be realized in an incomplete way because of the resistance natural language makes when it will accommodate itself to the regular pattern of the meter. This also implies that the natural prosody of language can be more or less perverted when it is squeezed into the metrical scheme. As a result a tension appears - often very fruitful - between the meter and the rhythm of language, [4, 5, 1]. Consequently, in order to investigate the rhythm in poetry reading, we need to use both methods from literary metrical analysis, starting out from syllables that form metrical feet, and methods from the analysis of the natural prosody of language with a segmentation into interstress intervals, in Swedish headed by stressed vowel onsets. These refer to as phonetic feet.

The major problems of the present study are the following issues:

1. To what extent is isorhythm maintained in reading?
2. How do pauses within and between lines maintain rhythmic continuity?
3. What are the characteristic features of trochaic and iambic patterns as they appear in reading? Are weak and strong syllables in an iambic foot (weak + strong) different from those in a trochaic foot (strong + weak)? If so, to what extent are these differences attributable to metrical grouping effects, and to what extent are they implicit in meter specific word and syllable structures?

2. EXPERIMENTAL PROCEDURE

We have approached the problems outlined above in three steps. One is through sequences of nonsense syllables representing prototype iambic and trochaic rhythmical patterns. The next step was to construct "lab poems" of strict iambic and trochaic meter, based on almost the same word material, enabling a minimal contrast in composition. Finally we turned to the study of traditional Swedish poetry. In such an analysis they were well acquainted with traditional Swedish poetry. The recorded material was subjected to our routine data bank processing, involving segmental analysis from spectrographic records. We accordingly measured durations of individual speech sounds, syllables, pauses and interstress intervals, the latter measured from the onset of a stressed vowel to the next stressed vowel. Mean interstress durations were calculated for feet not spanning a pause or otherwise marked syntactic boundary. We also measured interstress intervals spanning pauses and lining junctions to determine a possible rhythmic coherence with mean interstress durations.

In this study we have looked at stressed vowels, their pauses and syllable boundaries. A sequence of a stressed vowel and a nonstressed vowel is contained in both syllable sequence S+W. Average durations of such syllable based feet are not necessarily the same as average interstress intervals but usually serve as good approximations if averaged over a sufficiently long reading. In a study of the water sprit "Nacken" (The Water Sprite) by E.J. Stagnelius and the iambic "Kung Karl" (King Charles) by E. Tegnéi, the trochaic poem contains five stanzas, each of four lines of four feet each. Most of the pauses occurred at line and stanza junctions. Pause spanning interstress intervals formed a regular...

3. RHYTHMIC COHERENCE

We shall here report some of the main results. A more detailed account will be given in [2]. The issues of isorhythm and rhythmic coherence of pauses are illustrated in Fig. 1, which pertains to a three-line iambic "lab poem": "Nu väns vind drar fram med lust, till liv och glädje slang och dans och väcker Juva minness Bild." Each line contains four feet of the metrical structure W+S. However, for the purpose of bringing out a regularity of interstress intervals it was more relevant to consider stressed vowel onsets. The duration of each interstress interval is plotted vertically against the foot number with the foot text included below. Two versions are included, a normal reading and a scanned reading. One may observe that interstress intervals tend to vary in length with the number of associated phonemes. This is less so in the scanned reading in the normal reading mode the regularity of syllable sequences, implied by the meter, limits the variability of foot durations and may take some degree of isorhythm compared to prose reading. All the same we occasionally encounter large local variations of foot length. Pauses occurred at all line junctions of the "lab poem", Fig. 1. The interstress intervals, from the vowel onset of the last stressed vowel on the line to the onset of the first stressed vowel in the next line, are prolonged by about one mean interstress interval. The pause absorbs a silent beat. Accordingly, the line junction is divided into two parts of equal duration placed at the end of one line and at the beginning of the next line, which brings out the rhythmic continuation.

The same tendency of rhythmic coherence across pauses was observed in the reading of traditional Swedish verse and with greater consistency than in prose reading. Our most detailed data are from two poems, the trochaic "Nacken" (The Water Sprite) by E.J. Stagnelius and the iambic "Kung Karl" (King Charles) by E. Tegnéi. The trochaic poem contains five stanzas, each of four lines of four feet each. Most of the pauses occurred at line and stanza junctions. Pause spanning interstress intervals formed a regular...
lar pattern of preferred durations of approximately m=1,2,3,4 or 5 times a quantal module of To=525 ms. Out of the 21 occurrences, 12 were found on the m=2 level, 4 on the m=3 and m=4 levels, and one m=5. The quantal base, To=525 ms, is somewhat smaller than the average interstress interval, Ta=580 ms. We do not claim exact synchrony.

The iambic verse is made up of four line stanzas, normally with three complete (W+5) iambic feet per line with a regular occurrence of an extra weak syllable (hypotactic) at the end of each odd numbered line. Pauses were generally shorter after the odd lines than after even numbered lines, which tended to secure an overall regular timing of all line junction spacing interstress intervals to approximate 2Ta. In other words, at the end of the odd lines the pause acts as a supplement to the hyperstases, completing a foot. In the even numbered lines the pause alone adds a silent foot. The concept of rhythmical continuity across a pause or a line junction can be given two slightly different formulations. One is an invariance of a measure of pause duration plus the associated prepause final lengthening, which tends to amplify the measure of nTa. This seems to hold rather well for rhythmic reading of prose. In poetry, on the other hand, we found a more consistent trend of the entire spacing interstress intervals to comply with a measure of mTa.

This is what we could expect from a higher demand for rhythmic regularity in poetry reading. In a situation where the W/Ta and both models fit the data, we can expect m=m+1, i.e. one rhythmic unit is contained in the physical sound segments of the spanning foot.

4. METER SPECIFIC PATTERNS

Our next problem has to do with rhythmic patterns of read poetry in relation to metrical patterns. If the W of the iamb or trochee significantly from the W of the trochee or iamb, the two would be the case for the S of the iamb and the S of the trochee, the sole difference would be whether a line started with a strong or a weak foot. Also it is a line was terminated. It has been claimed in the literature, [6,7], that the S/W durational ratio is larger for iambic than for trochaic verse. This we have verified in the reading of our "lab poems" where an iambic version has the same text as the corresponding trochaic version with a weak initial syllable. However, we found that the reading of the strong syllable in the iamb, S=375 ms, was only slightly longer than the S=355 ms of the trochee. A great difference was found in the weak syllables with W=150 ms for the iamb compared to W=225 ms for the trochee. However, it is important to consider that this in part reflects differences in syllable complexity. The average number of phonemes per weak syllable was 2.55 for the trochee and 2.15 for the iamb. The average duration of a syllable is approx. 35% proportional to the number of phonemes, [3]. About half of the 75 ms difference between the W of the trochee and the W of the iamb, is attributable to the difference in syllable complexity, whilst the remaining 35 ms represents a true meter determined effect. The difference in strong syllable duration comparing the iamb and the trochee may entirely be explained by the slightly higher average number of phonemes in the iamb S than the trochee S. The main durational difference thus lies in the weak syllable, which is shorter in the iamb than in the trochee.

How do we explain these differences? First of all, the durational patterns we have found merely constitute one part of a complex pattern also carried by intonation and intensity contours that contribute to the lively character. The iambic reading compared to the more level trochaic reading. We may expect that the meter imposes a grouping effect in the read poem so as to enhance the final syllable of the foot, as a consequence of the trochee and the S in an iamb. This would essentially be the case of a terminal lengthening of the end of a line, which would enhance a trochaic W and an iambic S.

We have also looked into meter specific choice of language material. We have found a predominance of monosyllabic words in iambic as well as in anapestic poems, whilst trochees and dactyIes show a relative predominance of disyllabic words. It remains to be seen to what extent interstress patterns condition durational patterns in poetry reading.

We plan to compare our data above with the results of an earlier study of stressed syllable duration in connection with the number of phonemes contained. In these regression equations we insert the number of phonemes per W and S syllables of the read poetry we arrive at a S/W ratio of 2.4 for an iambic pattern and 1.9 for a trochaic pattern to be compared with the observed S/W=2.5 for the iambic verse and S/W=1.6 for the trochee verse. This shows that the projection of how language structure might impose constraints on poetry reading supports what we have already seen, that the contrast between the iambic W and the trochee W of the durational patterns is greater than implied by a language model derived from prose reading.

5. GENERAL DISCUSSION

We have dealt with two major aspects of temporal organization. One is the tendency of pause and line spanning interstress intervals to synchronize on a multiple of a basic rhythm module close to an average foot interval. In prose reading similar rhythmic traits were observed, but here the pause plus prepause lengthening is a more stable unit than the duration of the pause spanning interstress interval which varies with the number of phonemes contained. The other main problem concerns meter specific rhythmic patterns. The strong syllable duration is of about the same length in iambic and trochaic verse, whilst the weak syllable is significantly longer in trochaic verse than in iambic verse. One might suggest a smaller number of phonemes in the weak iambic syllable. The remaining difference could also in part be related to other meter specific selections of word and accent types. However, the durational patterns we have observed appear to reflect a specific grouping of a metrical foot with a poetic mode of reading, e.g. the relative liveliness of the iamb. In this respect we may claim that specific iambic and trochaic patterns are not "metrical myths" [6], but a reality as proposed by earlier investigators, [7,8].

A large number of problems remain to be tackled, e.g. the integration of other stress attributes such as stress and intensity variations into an overall model of poetic performance. Now, in the age of free verse these problems might seem antiquated. However, there is a recent trend in poetry writing of rediscovering the poetical virtues of metrical structures.

ACKNOWLEDGEMENTS

These studies have been supported by grants from The Bank of Sweden Teng- tercentenary Foundation, The Swedish Council for Research in the Humanities and Social Sciences and The Swedish Board for Technical Development.

REFERENCES

A CONTRASTIVE ANALYSIS OF SPANISH AND CATALAN RHYTHM

M. Carrió Font and A. Ríos Mestre

Departament de Filologia Espanyola - Laboratori de Fonètica
Universitat Autònoma de Barcelona, Spain.

ABSTRACT.

Durations of syllables, segments and pauses of similar texts in two languages, Catalan and Spanish, are compared at three speech rates: slow, normal and fast. These texts are read by one native speaker of each language. Results reveal that, although the temporal compression phenomenon has not exactly the same behaviour, both languages seem to be syllable-timed. Catalan tends to a proportional reduction in all syllables; Spanish tends to a more proportional duration of all syllables through stressed syllable reduction.

1. INTRODUCTION.

The aim of this paper is to study temporal compression of segments and syllables in Catalan and Castilian Spanish due to the influence of increasing speech rate. This temporal compression phenomenon is expected to be different in the two rhythmic categories traditionally classified: in stress-timed languages speech rate increase shows a higher degree of reduction in unstressed than in stressed syllables; in syllable-timed languages speech rate increase shows a proportional reduction in all syllables [1]. On the other hand, in syllable-timed languages, a ‘greater speed’ and an ‘easier articulation’ are achieved at the expense of consonants rather than vowels [3]. According to the classical literature about Spanish we have considered it to be a syllable-timed language [5]. Catalan has still not been studied from this perspective, although there are acoustic cues which indicate that it belongs to the same rhythmic category [2]. For this reason, a similar behaviour is supposed to occur in the temporal compression phenomenon.

2. PROCEDURE.

2.1. Corpus.

We have analyzed two versions, in Catalan and Spanish, of the same text: the fable “The North Wind and The Sun” (see Den Os [4] for the study of this same text in Dutch and Italian). It was read at three different speech rates, slow, normal and fast, by a native speaker of each language.

2.2. Subjects.

One native speaker of Catalan and one native speaker of Spanish acted as informants. Both were male and they were speakers of the standard variety of their languages. They had no difficulty at speaking at the requested speech rates.

2.3. Recording and acoustic analysis.

The speakers read the texts at three speech rates in one single recording session. It took place in a sound isolated room in semi-anechoic conditions at the Phonetics Laboratory at the Universitat Autònoma de Barcelona. A Sennheiser MD 441N directional cardioid microphone and a Revox A77 tape recorder were used.

The signal was digitized at 10 KHz sampling rate using the routines implemented in the MacSpeech Lab II software package by GW Instruments running on an Apple Macintosh II.

The audio wave was segmented and durations were measured on the oscillographic representation, locating the boundaries of sounds using changes in the waveform as the main criteria. When necessary, spectrographic and perceptual checking listening to the segments were also used.

3. RESULTS.

The Catalan text contains 171 linguistic syllables and Spanish one contains 179. The overall time of readings (included pauses) of the Catalan version is 39.1 s. (slow), 32.4 s. (normal) and 25.8 s. (fast) and of the Spanish version 37.1 s. (slow), 32.9 s. (normal) and 26.4 s. (fast). This means that the overall speech rate -expressed in linguistic syllables per second- in Catalan readings is 4.4 (slow), 5.3 (normal) and 6.6 (fast) and in Spanish readings is 4.8 (slow), 5.4 (normal) and 6.8 (fast). The number of syllables per time unit seems to be a good objective measure of speech rate. The versions of the languages may be compared with respect to speech rate. Values for each speech rate in the two languages are similar and there is an inversely proportional relation between speech rate increase and total duration decrease as expected.

The overall time of pauses in Catalan is 5.7 s. (slow), 5.3 s. (normal) and 3.6 s. (fast); and in Spanish is 8.9 s. (slow), 6.5 s. (normal) and 3.7 s. (fast). Values are higher in Spanish than in Catalan except in the fast reading, in which they are practically the same. The articulatory rate (excluding pauses) in Catalan is 33.4 s. (slow), 27.1 s. (normal) and 22.2 s. (fast) and in Spanish is 28.2 s. (slow), 26.4 s. (normal) and 22.7 s. (fast). The articulatory rate -expressed in linguistic syllables per second- in Catalan is 5.1 (slow), 6.3 (normal) and 7.7 (fast) and in Spanish 6.3 (slow), 6.8 (normal) and 7.6 (fast). We observe that the articulatory rate increase in Catalan is proportional in the three readings, but in Spanish there is a weak increase between slow and normal readings, and a more noticeable increase between normal and fast readings. Anyway, articulatory rates corresponding to slow and normal readings have a higher value in Spanish than in Catalan, although the differences between articulatory rate values decrease; and, finally, the values for fast reading in both languages tend to be similar.

The number of syllables realized in Catalan readings is 166 (slow) and 165 (normal and fast), and in Spanish readings it is 178 (slow and normal) and 177 (fast). The overall speech rate -expressed in phonetic syllables per second- in Catalan readings is 4.3 (slow), 5.1 (normal) and 6.4 (fast), which are perfectly comparables with Spanish values: 4.8 (slow), 5.4 (normal) and 6.7 (fast). Articulatory speech rate expressed in phonetic syllables shows the same behaviour that the rate expressed in linguistic syllables, but the fast reading value is not so similar between both languages. Those values for Catalan are 5.0 (slow), 6.1 (normal) and 7.4 (fast) and for Spanish 6.3 (slow), 6.7 (normal) and 7.8 (fast).

It is then clear that there are some problems connected with expressing speech rate in syllables per second. Questions arise as to whether pause-time has to be included and which types of syllables have to be counted, phonetic or linguistic ones. We have computed the overall values of linguistic and phonetic syllables, and of speech and articulatory rate. But we have taken only into account values of phonetic syllables because they correspond to the actual phonetic realization of the segments, for the same reason, values corresponding to speech rate have been used, because among other factors, it is no possible to distinguish pauses from stop gaps occurring after a pause. Then if we take only into account articulatory rate values, some information would be lost.

In order to study the temporal compression phenomenon as a function of the speech rate increase, regression analysis has been applied taking into account the following conditions for three speech rates in both languages: (a) the overall speech rate, expressed in phonetic syllables per second as an independent variable. (b) as dependent variable, in each case: the mean duration of unstressed syllables, stressed syllables, vowels, stressed vowels, unstressed vowels, Catalan schwa, consonants, obstruents, and sonorants.

The relative decrease in duration per syll/s is the following:
3.1. Syllables. Catalan unstressed and stressed syllables show an analogous shortening, which is higher in the stressed than in the unstressed ones (30.6 vs. 25.0). Spanish stressed syllables shorten to a lesser extend considering the behavior Catalan syllables (20.4), and Spanish unstressed syllables present an even lower degree of shortening (7.2). See Figure 1.

3.3. Stressed vowels vs. unstressed vowels. Differences in shortening between stressed and unstressed vowels in both languages are clear, although they are more prominent in Spanish (10.6 and 4.3) than in Catalan (18.2 and 9.0, respectively). Catalan has a schwa, which undergoes a shortening similar to the overall unstressed syllables (10.4). See Figure 3.

3.2. Vowels vs. Consonants. Vowels and consonants are subject to a similar shortening in both languages; however it is lower in Spanish (6.2 and 4.2) than in Catalan (12.6 and 8.7, respectively). See Figure 2.

3.4. Obstruents vs. sonorants. In Spanish there is a great difference in the shortening between both types of consonant categories (6.1 and 1.5). In Catalan, in which the degree of shortening is higher between sonorants and obstruents are not so important (9.6 and 7.0, respectively). See Figure 4.

Figure 1: speech rate vs. syllable duration.

Figure 2: speech rate vs. vowel and consonant duration.

Figure 3: speech rate vs. vowel duration.

Figure 4: speech rate vs. consonant and sonorant duration.

4. DISCUSSION.

All categories studied show a higher degree of shortening in Catalan than in Spanish. Considering that in Spanish the three speech rates are a bit higher and the shortening is a bit lower than in Catalan, we can expect that temporal compression as a function of the speech rate increase would be smaller in Spanish than in Catalan.

On the other hand, considering that stressed syllables have the longest duration in Spanish, the fact that they are subject to a higher degree of shortening than unstressed ones reveals a strong tendency towards equal syllable duration. The same phenomenon is found for vowels in both languages. This seems to imply that Spanish and Catalan tend to syllable-timed languages.

The ratio between the degree of reduction of vowels vs. consonants is the same in both languages (1.5). Temporal compression of vowels is higher than of consonants. The behaviour of those syllable types in Spanish seem to be in disagreement with Dauer conclusions [3]. However, we believe this behaviour is coherent with the results obtained in our experiment, which reveal that the categories of syllables and segments with longer mean duration are shortened in a higher degree. According to Bertinetto [1], we can conclude that Catalan and Spanish are not stress-timed languages, because speech rate increase does not show a higher degree of reduction in unstressed than in stressed syllables. They would be then considered syllable-timed languages.

Speech rate increasing in Catalan shows a proportional reduction in all syllables. Speech rate increasing in Spanish shows a higher reduction in stressed syllables than in unstressed ones, although stressed syllables are always the longest ones. Then, there is a tendency to shorten longer segments and stressed syllables most. Through stressed syllable reduction, proportional duration of syllables tends to be achieved.

5. CONCLUSION.

It has been shown that both languages tend to be syllable-timed, although the processes involved are not exactly the same. The fact that Spanish seems to make equal syllable durations (through stressed syllable shortening related to speech rate increase) suggests that its rhythm is syllable-timed as it has been traditionally defined; syllables recur at regular intervals. In Catalan the temporal compression is more marked almost equally in all syllables, so that we can presume that its rhythm is syllable-timed in agreement with Bertinetto’s proposal [1].

However, in order to characterize a language from a rhythmic point of view there are other factors to be taken into account. Furthermore, we are aware of problems concerned with our experiment:
- segmental reduction is also constrained by syllable structure, segment position in the utterance or speech style.
- the fact that reading rates are constrained affects the degree of naturalness of the corpus.
- the preliminary results of this study suggest that more research is still needed in order to describe accurately the temporal compression phenomenon.

6. REFERENCES.

RHYTHMICAL MODEL OF A PHONETICAL WORD OF PRESENT-DAY LITHUANIAN UTTERANCES

L. Anusiene

Technical University, Vilnius, Lithuania.

ABSTRACT

The aim of this investigation was to study duration as a prosodic component of the rhythmical structure of phonetical words of different accent type in utterances typical of Standard Lithuanian and to discover the temporal characteristics of a rhythmical model of a phonetic word, taking into account the inherent prosody of vocalic and diphthongal syllabic nuclei. The obtained model has revealed the main regularities in the distribution of duration of stressed and unstressed syllabic nuclei.

1. INTRODUCTION

In a previous study /1/, an attempt was made to investigate durational characteristics of acute and circumflex vowels and vocalic diphthongs in extended speech contexts. The distinguishing feature of Lithuanian stress is that homogeneous long monophthongs and vocalic or mixed diphthongs may have acute or circumflex accent. Having experimentally proved that there is no significant difference in duration neither between acute and circumflex nor diphthongs in extended speech, we analysed vocalic and diphthongal syllabic nuclei irrespective of the accent type.

The experiment reported below is an extension of the previous investigation: this time involving durational ratio of stressed and unstressed syllabic nuclei of all types.

2. THE EXPERIMENT

The experiment corpus consisted of 128 utterances, recorded by 3 male and 2 female subjects. Measurements were obtained from intonograms. In the experimental material the vowels and diphthongs under investigation were presented in different phonetical environments and in various positions in the phrase. So as to compensate for the influence of word position in the utterance, they were constructed so that the vowel was found an equal number of times in each position. In order to compensate for differences in absolute duration in different positions, computations were based on relative differences in duration. The data for each subject were individually analysed, but since the same corpus was used for each subject we also contrasted the data on vowel and diphthong duration for all the subjects as a group.

Since two- and three-syllabic words make up the most recurrent accentual pattern in the Lithuanian language, the temporal characteristics of rhythmical structure of such phonetical words have been investigated.

3. RESULTS

Certain durational distribution of stressed and unstressed syllabic nuclei make up the main feature of the Lithuanian rhythm.

The analysis of durational ratio of stressed and unstressed short vowels of the same height revealed that:

a) there is almost no difference in the length reduction of the lst pretonic long vowels /a/ and /a/ (0.77:1 and 0.79:1 respectively);

b) there is a great difference in the length reduction of the lst pretonic high vowels /u/ and /u/ (0.67:1 and 0.87:1 respectively);

c) the length reduction of the 2nd pretonic vowels is greater than that of the lst pretonic vowels, /u/ being subjected to the highest degree of reduction (/a/, /a/, /u/, /u/ — 0.74:1, 0.76:1, 0.70:1, 0.70:1 respectively);

d) the length reduction of the 2nd posttonic vowels is weaker than that of the lst posttonic vowels (/a/, /a/, /u/, /u/ — 0.74:1:0.86, 0.76:1:0.80, 0.70:1:0.83, 0.81:1:0.98 respectively);

e) the length reduction of the 2nd posttonic vowel is rather small, with /u/ being even longer in duration than the stressed one (/a/, /a/, /u/, /u/ — 1:0.94, 1:0.85, 1:0.92, 1:1.1 respectively).

The analysis of durational ratio of stressed and unstressed long vowels of the same height revealed that:

a) the length reduction of the lst pretonic long vowels is very similar to that of short vowels, with /u/ being subjected to the highest degree of reduction. Long vowels /a/ and /u/ were not included into the experimental material as they are very rare in the pretonic position in the Lithuanian language (/a/, /a/, /u/, /u/ — 0.82:1, 0.70:1, 0.58:1, 0.70:1 respectively);

b) the 2nd pretonic long vowels like the short vowels have a tendency to a greater length reduction;

c) the 2nd posttonic long vowels have a tendency to a greater length reduction than the short vowels.

In a previous study /1/ it was revealed that there is essentially no difference in duration between the circumflex and acute diphthongs /ei/, /i/ and /ou/, while there is statistically significant difference in duration between diphthongs /ai/ and /au/ pronounced with different accent type.

The analysis of durational ratio of stressed and unstressed diphthongs irrespective of the accent type revealed that:

a) the diphthong /ai/ has a greater length reduction in the lst pretonic syllable than in the lst posttonic syllable as in short and long vowels (0.74:1);

b) the diphthong /e/ contrary to the diphthong /ai/ has a greater length reduction in the lst posttonic syllable than in the
The analysis of the diphthongs /ai/ and /au/ pronounced with different accent type revealed that:
a) the acute and circumflex diphthong /au/ has greater length reduction in the 1st posttonic syllable than in the 1st pretonic syllable (/āu/, /āu/ —— 0.68:1:0.66, 0.80:1:0.78 respectively);
b) contrary to the diphthong /au/ the acute and circumflex diphthong /ai/ has a greater length reduction in the 1st pretonic syllable than in the 1st posttonic syllable (/āi/, /āi/ —— 0.62:1:0.64, 0.77:1:0.80 respectively).

The analysis of durational ratio of stressed and unstressed vocalic and diphthongal syllabic nuclei revealed the temporal characteristics of a rhythmic model of a phonetical word. According to this model, the following regularities in the distribution of stressed and unstressed syllabic nuclei may be distinguished:

1. The length of unstressed syllables is dependent on the distance from the stressed syllable, with syllables closer to the stress being longer.
2. The pretonic syllables show greater reduction in duration than the posttonic syllables.
3. The 1st pretonic syllable is approximately equal in length to the 2nd posttonic syllable.
4. The 2nd posttonic syllable is approximately equal in length to the 1st posttonic syllable.

It is assumed /2, 3/, that posttonic syllables word or phrase finally are longer than pretonic syllables. It is conditioned by syllable to the stress position as well as by intonation. It remains to be proved, however, whether the above described temporal structure is language specific or language universal.

4. REFERENCES

Incidence du trait phonologique de durée vocale sur la prosodie du français québécois

Laurent Santerre
Département de linguistique
Université de Montréal (Québec) H3C 3J7

Abstract
The main prosodic differences between Quebec French and French from France originate in the vocalic system. Quebec French retains the old system of long and short vowels in which the distinctive feature of duration is imposed on morphology, independently of degree of stress and syllabic derivation. This durational contrast creates rhythm that excludes syllabic isochronism. In Quebec French, in an intonational stretch of two consecutive syllables, both syllables can be optionally stressed by using different means (intonation and duration) apparently freely distributed.

Introduction
On peut penser que les règles prosodiques liées à la syntaxe et à la sémantique ont des chances d'être communes aux différents dialectes français, tandis que celles qui sont longues plus étroitement par la phonologie, la phonétique et la pragmatique sont plus spécifiques; c'est le cas pour ce qui est du français québécois. Par contre, les règles prosodico-syntaxiques, prosodico-sémantiques et rythmiques que Mario Rossi (1985 et 1987) a formulées à partir d'exemples de français de Paris, sont communes aux français des deux côtés de l'Atlantique.

Les règles sont assez générales pour avoir un statut phonologique aux frontières des principaux constituants syntaxiques. Les incréments continuitatifs ou conclusifs se retrouvent aux mêmes frontières du Québec et en France; c'est le cas des /<t/, /<t/, /<t/, /<t/ et /<t/, même s'ils se réalisent pas nécessairement en surface phonétique par les mêmes variations paramétriques. Les règles rythmiques et d'ajustement peuvent rendre compte des nombreuses variations phonétiques liées au style, au débat, avec une certaine liberté laissée à la spontanéité des locuteurs.

Particularités prosodiques
Les particularités prosodiques que je signale ici tiennent au système phonologique des voyelles longues et brèves que les Québécois ont en bonne partie hérité de l'ancien système vocalique du français de l'Île-de-France. La durée phonologique omniprésente dans le français québécois a des incidences sur les modes de dénouements phonétiques de l'accentuation, le système prosodique, sur l'organisation temporelle à l'intérieur de la syllabe et du mot, sur la rythmique non isochronique de la phrase et sur le placement des accents secondaires dans l'enoncé. Je me limiterai ici au rôle de la durée dans l'accentuation.

Voyelles longues et voyelles brèves
Des 17 voyelles phonologiques du français québécois, huit sont longues par nature et neuf le sont pas. Ces longues sont : /a/ de fête, opposé à la brève correspondante /a/ de faible; /a/ de pâte, opposé à la correspondante /a/ de patte; /e/ de côte, opposé à la brève /e/ de cote; /e/ de jeune, opposé à la brève /e/ de jeun; dans ce groupe de voyelles orales, l'opposition de durée s'ajoute à l'opposition de timbre et peut être neutralisée. Les quatre voyelles nasales sont aussi longues par nature (Santerre 1974).

Cos huit voyelles longues par nature s'allongent peu par coarticulation avec les constrictives sonores qui les entrent et d'autres se laissent peu abrégés par les oucussives sonores (Santerre 1987) [2]. Les voyelles brèves sont indifférentes au trait phonologique de durée, mais elles sont considérablement allongées et abrégées par coarticulation consonantique; ce sont les voyelles hautes /i, y, u/ et les quatre brèves /e, a, o, e/ opposées aux quatre longues orales; deux voyelles, le /e/ et le /a/, ne se trouvent pas en syllabe entrelée.

La rencontre dans la rime des sept voyelles brèves avec les codas allongées, ou brèges, ou indifférentes (oucussives sonores et constrictives sonores) occupe la production de trois groupes de syllabes caractérisées par leur durée spécifique (Santerre 1987) [1]. De même, la rencontre dans la rime des huit voyelles longues par nature avec les trois classes de consonnes engendrent des groupes de syllabes plus ou moins longues.

Les rapports de durée
Le rapport de durée des voyelles brèves et des voyelles longues est considéré même en québécois. Toutes choses égales d'ailleurs, il peut varier de 1,5 à 3 et même beaucoup plus, parce que les voyelles hautes en dehors de l'accent peuvent être syncopées ou très abrégées, elles ne représentent qu'une faible fraction de la durée d'une longue; ainsi le [k] de compté peut se faire de 0 à 5 ou 6 cs, tandis que le [k] de commenter peut faire 12 à 20 cs. Ces rapports ne sont pas qualitatifs. Les voyelles hautes, en s'exaspérant ou en se syncopant en dehors de l'accent, abrègent et même fon disparaître une syllabe, ce qui oblige les syllabes voisines à s'allonger pour intégrer les consonnes laissées sans voyau vocalique (Archambault 1985, J.-F. Couturier, recherche de doctorat en cours).

Durée morphologique lexicaux
Les syllabes à noyau long par nature qui constituent des morphèmes lexicaux fréquents gardent leur durée vocale caractéristique même quand elles entrent en composition avec d'autres syllabes pour former des lexèmes; et dans ce cas, la coupe morphologique peut avoir priorité sur la coupe syllabique dans la prononciation. Exemple : les morphèmes longs /tête/ et /pâte/ se prononcent en respectant la durée et l'entraîne dans : tête à l'ouvrage /tête/ et /pâte/ à l'ouvrage /pâte/. Il est à remarquer que les morphèmes à noyau bref allongé par coarticulation n'ont pas cette priorité de la coupe morphologique sur la coupe syllabique. Exemple : /sages/ /sagé/ a un noyau allongé qu'on ne trouve pas dans /sagesse/ /sagésse/ à cause de la dérivation syllabique, mais qu'on retrouve dans /sagement/ /sagement/.

La durée dans la morphologie verbale
À la faveur des fusions vocaliques qui mettent en cause les flexions verbales, les contractions syllabiques suivent les marques morphologiques de temps au moyen de la durée distincte. Dans un test au moyen de phrases synthétisées, j'ai fait varier la durée vocale dans un mot entre deux phases [5]: "Il est à Pau". Une certaine d'étudiants québécois ont perçu le présent quand le /a/ était bref, et l'imparfait quand il était long. L'explication réside dans la durée qui représente la fusion des deux voyelles sous-jacentes de l' *êté à Pau*; je *te* *pat*; je *te* *pat*; je *te* *pat*; je *te* *pat*; je *te* *pat*. Il faut faire le passage au niveau phonologique la représentation des deux voyelles sous-jacentes de l' *êté à Pau* /te* *pat*; je *te* *pat*; je *te* *pat*; je *te* *pat*; je *te* *pat*; je *te* *pat*; je *te* *pat*; je *te* *pat*. Ce test a été réussi presque sans exception par les Québécois, et on n'a pas d'échecs. Les phrases suivantes ont été complètement
confondues par quinze auditeurs québécois; lues par un locuteur montréalais, elles ont été distinguiées à 77%. Les mesures montrent que la durée des syllabes morphologiques auto-
ombres sont significativement plus longues en québécois. Il ne s'agit pas d'un allongement accentuel, mais d'une trace de la durée liée aux articles contractés (des = de les). Dell (1984, p. 100) dit qu'il ne semble pas qu'on puisse jamais marquer une opposition de longueur en syllabe inaccentuée. C'est sans doute le cas en français de Paris; en québécois la durée garde encore souvent sa pertinence même en dehors de l'accent.

Ces considérations ont pour but de bien établir le fondement phonologique et morphologique de la durée en québécois, durée qui a une incidence considérable sur la procédure. La durée en français de Paris n'a pas ce statut fondamental; elle est seulement physiquement conditionnée par l'accentuation et par la coarticulation consonantique. Elle ne met pas en œuvre comme en québécois une commande phonologique de production et de détection qui renforce l'effet mécanique involontaire.

**Incidence de la durée sous-jacente sur l'accentuation**

Je prendrai mes exemples dans l'intéressant article de Dell (1984). Les intuitions phonologiques de l'auteur sont illustrées par une centaine de phrases que je lui ai demandé d'enregistrer en studio. Un certain nombre de ces phrases ont été soumises à des tests de perception auprès d'auditeurs, aussi bien français que québécois; elles ont été différemment distinguées par les uns et par les autres. L'analyse prosodique instrumentale et psychacoustique rend bien compte des cas de confusion: l'accentuation de Dell a été réalisée dans ces enregistrements presque exclusivement par l'intonation.

Pour des raisons d'eurythmie, Dell déplace l'accent 2 dans (a) et (b):

(a) La faîte sert à faucher l'oselle

```
0 2 0 0 0 0 0 1 (2-6)
```

(b) La faîte sert à faucher l'oselle

```
0 0 0 0 0 0 0 3 (5)
```

Comme le prévoit l'auteur, (b) devient homophone de (c): "la faîsaient à faucher l'oselle".

Dans un test de compréhension auprès de seize Québécois étudiants de phonétique, (a) a été entendu comme la faîte par tous, (b) ne l'a été que par un seul. L'explication se trouve dans le fait que Dell (p. 88) est obligé de désaccentuer la faîte parce qu'elle accentuée la syllabe suivante celles. La même phrase prononcée par des Québécois, qui dépaissent aussi l'accent sur ce l, ne change pas de sens, parce que faîte conserve une durée qui sauve son statut de synchronie nominal sujet. L'accent de faîte est fait par la durée et celui de ce l est fait par l'intonation. Selon Dell, une règle de non-contiguité accentuelle dans un même tronçon prosodique entend d'accentuer en français deux syllabes consécutives. C'est sans doute parce que le lézard n'a pas le temps de faire les ajustements nécessaires pour réaliser deux intonèmes distincts sur des voyelles voisines. Dans un dialecte qui table aussi bien sur la durée que sur le Fo pour faire l'accentuation, rien ne perd le locuteur de faire deux accents consécutifs pourvus qu'ils soient réalisés par des paramètres différents (Santerre 1990) [2].

**La règle d'allongement de Dell**

Dell (p. 101) reconnaît que des phrases homophones comme (a) et (b) peuvent être distinguées par l'allongement d'une syllabe accentuée. (a) Des des odorants; des odorants, 0 2 0 0 1 (a) (d) des déodorants; des déodorants, 0 0 0 0 1

Cette règle stipule qu'on allonge facultativement la syllabe finale d'un mot accentuable, mais non pas la syllabe précédente des mots féminins qui est suivie d'un muet, comme par exemple: C'est pourquoi Dell allonge l'accent secondaire dans (c) et dans (d): (c) ce - lui qui par le coud (d) ce - lui qui par le coud

```
0 3 0 2 0 1
Fo: 124 156 (sourd) 139 105
Durée 136 211 150 369 240
```

```
0 3 0 2 0 1
Fo: 127 151 (sourd) 140 111 110
Durée 150 209 132 303 129 224
```

Remarque: ici ce n'est pas l'écart du Fo sur la syllabe accentuée qui fait remarquer l'intonation sur 2, mais le long glissando vers la syllabe suivante:

"On peut dire en québécois, sans se soucier de la continuité accentuelle: Celui qui part coud 0 3 0 2 1" Dans la lecture de la phrase suivante, aucun locuteur québécois n'a fait entendre ruoucou, comme le fait Dell: "Les seuls de l'élève couleu".

**Conclusion**

Si une relative isochronie syllabique dans le langage des Parisiens peut être contestée, à combien plus forte raison se trouve-t-elle exclue de celui des Québécois. En France, la durée phonologique comme trait distinctif des voyelles est perdue, même si on peut encore en entendre des traces. Le français québécois, au contraire, est obligé d'organiser temporellement complexe des syllabes pour respecter les durées imposées par le système phonologique. Sa démarche rythmique se rapproche de l'anglais américain dont le système vocalique exploite l'opposition de durée et de timbre.

On comprend facilement que le trait de durée, qui est incontournable aux niveaux phonologique et phonétique, conserve ses droits jusque dans la morphologie et s'impose dans le rythme des énoncés et dans les formes de l'accentuation en québécois.

**Références**


PERCEIVING RHYTHM IN FRENCH?

Jacqueline Vaisseiré
CNRS URA-1027
Institut de Phonétique
19, rue des Bernardins, 75005 Paris, France

ABSTRACT
This paper deals with the problem of rhythm in French, a language with no strong stress contrast. First, oversimplified patterns at three different levels, the breath group (BG), the prosodic word (PW) and the CV syllable (CV) are proposed as archetypical reference rhythmic patterns. These 3 layers seem to correspond to psychological realities. BG layer, the larger one, consists in the alternation between 2 highly contrastive global groups. PW layer is characterized by the repetition of variants of an archetypal word pattern, shaped by at least one oscillation of pitch between the high and low registers, with a durational peak on its last sounded syllable, marking its end. The last layer is the succession of typically rising, tense CV syllables with soft onset. One of the three layers may become perceptual more dominant than the others for the perception of rhythm, depending on the speaking mode. Second, despite important differences, PW layer in French and the tone group in English are interpreted as 2 variants of the same archetypal psychological pattern where accentuation and lengthening are associated with the notion of beginning and end, respectively. In English, accentuation is dominant, and lengthening recessive. In French, it is the contrary, but accentuation is also intrinsically present (emphatic stress and initial rise at word beginning) leading to some combination in the present-day scheme of French rhythm.

INTRODUCTION
In speech, the notion of rhythm is often based on the perception of stress and recurring prominent syllables. Heffner notes that "languages with strong stress are likely to have rhythms of no subtlety whatever; languages which make less use of stress contrast have rhythms which are less obvious." (Heffner, 1950:227). Naive speakers of French do not have a clear idea of what a "stress" can be, and locating prominent syllables in non emphatic French is a difficult task. What about rhythm in French, which obviously is not primarily based on the perception of an alternation between stressed and unstressed syllables?

1. THE MULTILAYERED TEMPORAL RHYTHM
There seems to exist 3 perceptual units which give rise to a multilayered rhythm in French: (i) two basic global tunes, (ii) an archetypal "prosodic word" (PW) pattern; (iii) a typical open syllables CV. It is difficult to disentangle the different units in a purely acoustic study since the 3 layers interfere. The following caricatural patterns should be interpreted as prototypical percepts toward which the acoustic realizations tend to correspond (see Figure 1).

1.1: The two BGs
The first ingredient of the alternation of two highly contrastive global contours at the level of the breath group. The contrast between BG+, ending by a sharp rise on the final syllable and BG-, ending by a sharp fall extend over several syllables and seems largely "exaggerated" in French, as compared to English (Delattre, 1966:75). Both BGs are characterized by final lengthening.

1.2: The PWs

degree of dependency of the PW with the following PW (the more rising, the more independent). Rhythmic constraints play also role in the choice of PW, each individual speaker tends also to repeat the same PW (Vaisseiré, 1974:255).

Both the duration and melodic profiles may be described as "rising", from short and low for the function word at the PW beginning to high and long syllables at the PW end (see also Delattre, 1966; Touati, 1987), with a plateau on the intermediate syllable (s).

Figure 1 only represents main tendencies observed in data, and rather correspond to hypothesized archetypical concepts. The linguistic realisation of the PW is obviously disturbed by a number of conflicting influences: (i) a short-long alternation (Duez et al., 1984); (ii) longer duration of "heavy" syllables (closed syllables and syllables with nasal vowels) and end of morphemes; (iii) relative lengthening of the penultimate syllable as a mark of several regional "accents" (Carton, 1967); (iv) intrinsic and contrastive character of accent (as observed in other languages, Di Cristo & Hirst, 1986), and same vowels, such as /e/ are particularly short, even in final position. Nevertheless, the deviations of the duration and formant profile from the idealised curves seem to be most of the time explicable.

The PW notion corresponds to the traditional notion of "sense group". In terms of size, the PW corresponds to the stress group level in English. Because it does not have a clear anchor point such as a stressed syllable, the PW is probably less salient as a perceptual unit than the tone group, leaving more room for the syllable to play perceptually a more dominant role than in English.

L3: The CV syllables
The well-noticed perceptual saliency of the syllable as a rhythmic unit in French (Dauer, 1983, Wink & Wieland, 1982) is probably due to the lack of clear strong beat at the PW level, leading to an apparent uniformity of the syllables. Phonation is also perceived as uniformly particularly tense (no affricates, no lax vowels, not much reduction, no diphthongs and no diphthongized vowels, Delattre, 1966:323). Each syllable seems predominantly
open and "rising", with the vocal tract opening progressively up to the very end of the syllable, which typically ends with a vowel, with a delayed Fo peak and internal intonation (see Delattre, 1966:151), and a strong anticipatory coarticulation effect during the consonant preceding each vowel (Delattre, 1966:122), contributing to a softer attack (onset) of the vowel, as compared to English. The number of open syllables prevails in French (76% according to Delattre, 1965:42) and most of the syllables have the simple structure CV (54%-%). Since the simple CV structure is highly repetitive, it is a good candidate to become a pregnant percept (cf. the notion of "pregnancy" in the Gestalt Theory). PW and CV percepts coexist, such as the tendencies of giving same length to both the successive PW and the successive CV.

One of the 3 layers may be made perceptually more emergent than the others: isochronous syllables, in carefully spoken speech; same size PW in poetry, and regular BG in rapid, conversational speech. Interspeaker variability may be explained by the fact that each speaker is free to give more weight to one of the 3 main tendencies.

It is difficult to "prove" in a scientific way the coexistence of the different perceivable layers of simultaneity. The "pregnant" speech patterns stored in speakers' memories are often said to influence the way they perceive the different languages. Delattre's example of the trichotomy of a sentence in a given language by native of different languages (1965:23) seem to indicate that the stored basic patterns are very different (and not necessarily shared) for French and English listeners. Results of psychoacoustic experiments on the perception of rhythm in non-speech stimuli however reveal that French and English archetypical speech patterns, apparently very different, may be 2 variants of the realisation of a universal pattern.

2: TEMPORAL VERSUS INTENSIVE RHYTHMIZATION

Phonetic experiments on tone bursts have largely confirmed the role of a longer interval or of a elongation of a pulse as a right boundary marker, the role of accent, either intonational or pitch (a left boundary marker. They have shown a clear tendency to perceive the elements inside a grouping (once they have been perceived as grouped) as more isochronous than they actually are (Fraisse, 1956 and 1974 for a summary and references at Allen, 1966:151), and a strong anticipatory coarticulation effect during the consonant preceding each vowel (Delattre, 1966:122), contributing to a softer attack (onset) of the vowel, as compared to English. The number of open syllables prevails in French (76% according to Delattre, 1965:42) and most of the syllables have the simple structure CV (54%-%). Since the simple CV structure is highly repetitive, it is a good candidate to become a pregnant percept (cf. the notion of "pregnancy" in the Gestalt Theory). PW and CV percepts coexist, such as the tendencies of giving same length to both the successive PW and the successive CV.

One of the 3 layers may be made perceptually more emergent than the others: isochronous syllables, in carefully spoken speech; same size PW in poetry, and regular BG in rapid, conversational speech. Interspeaker variability may be explained by the fact that each speaker is free to give more weight to one of the 3 main tendencies.

It is difficult to "prove" in a scientific way the coexistence of the different perceivable layers of simultaneity. The "pregnant" speech patterns stored in speakers' memories are often said to influence the way they perceive the different languages. Delattre's example of the trichotomy of a sentence in a given language by native of different languages (1965:23) seem to indicate that the stored basic patterns are very different (and not necessarily shared) for French and English listeners. Results of psychoacoustic experiments on the perception of rhythm in non-speech stimuli however reveal that French and English archetypical speech patterns, apparently very different, may be 2 variants of the realisation of a universal pattern.

Prominence on final syllables was generally considered to be the rule in non emphatic French. There is however a large numbers of papers starting in the previous century which question this traditional point of view (see Fonagy, 1980, for a review). Fonagy & Fonagy (1976) have shown that when in conversational speech and story telling, final syllables were perceived as more prominent, in a journalistic style, initial syllables were perceived as more prominent in 74% of the cases. The widespread use of emphatic stress at the word beginning by the journalists and the politicians is less and less perceived as emphatic, but as a special style. The present-day French prosodic system is in the process of a change and the difficulty of present-day phoneticians on making firm statements on French prosody may be the expression of the on-going change. As a consequence, it is very difficult to make clear statements on French prosody, since there are typically at least two different prosodies.

CONCLUDING REMARKS

The French PW and the English tone groups may be interpreted as 2 variants of the same archetypical psychological pattern which associated accentuation with the beginning and lengthening of the word. In French, the accentuated element is more intense and lengthening recessive. In French, temporal organisation is predominant, but (initial) accentuation is also intrinsically part of the linguistic structure, and both making the study of rhythm a very difficult matter. Progress may come from experiments in non speech stimuli and from the investigation of prosody of other languages. The basic psychological constraints are integrated into the prosody and rhythm of diverse languages.

REFERENCES


FONAGY, J. (1980a), "L'accent français accent probabilistic dynamique d'un changement prosodique", in L'ACCENT EN FRANCAIS, CONTEMPOAIRI, Studia Phonetica 15, Fonagy, L. & Léon, P. (Eds.)
HEFFNER, R.M., (1950), GENERAL PHOTETICS, The Univ. of Wisconsin Press.
ABSTRACT
A model of tone production in Standard Chinese is presented and confronted to phonetic and EMG data. The model is of the command-response type; F0 is viewed as the response of the laryngeal structure to excitation commands. The same speech material was used to obtain EMG data and to model F0 contours so that tone production can be viewed from two different perspectives. EMG data reveal stable patterns of laryngeal muscle activity attached to each tone. Similar patterns obtain for the model commands.

1. INTRODUCTION
Second order linear systems are widely used for approximating the behaviour of various physical or biological systems, with respect to some given dimension. Indeed, one reason for this is that linear systems and their mathematics are well understood. For many natural systems however, biomechanical properties make the approximation by a linear system reasonable. For the laryngeal structure, Ohman [8] proposed the first to model F0 contours by the response of a linear system to excitation commands. Fujisaki [2] took up the idea for speech and singing voices, and proposed a biomechanical interpretation of his model [3]. In short, commands can be understood as controlling the length of the vocal folds, hence their longitudinal tension, hence F0 (for a more accurate account, see Bod [1]). However close to the physics these models may come, they would not be of much use if linguistically relevant patterns of commands could not be identified among the stream of the many commands required for modelled contours to closely follow actual Fo data. Ohman and Fujisaki both identified simple patterns attached to the production of pitch accent. We applied the same ideas to model F0 contours in Standard Chinese, proposing that qualitatively stable patterns of commands are attached to each tone type. Starting with the simplest patterns [9], we gradually came to the patterns presented in section 4. Our model not only provides an economical account of tone production, but also explain tone contour changes due to the tonal coarticulation that occurs in running speech. EMG studies of laryngeal muscles have evidenced stable patterns of activity attached to each tone type [5]. It is tempting then, to bring together Fo modelling and EMG data obtained with the same speech material.

2. MATERIAL
The speech material was designed for EMG experiments, where Cricothyroid (CT) and the Sternohyoid (SH) were examined. We used target syllables embedded in a frame sentence: /yi:ge X x4/ (a character X). Target syllables X belonged to minimal series sharing the same segmentals in the four tones: [i], [i], [i], and [u] (in Pinyin transcription, /i/, /i/, /i/, and /u/). Those segmentals were chosen in order to minimize SH contribution to supralaryngeal articulation (some SH activity related to tongue backing was expected for /u/). The target syllable X does not occur in prepausal position, is stressed and surrounded by unstressed syllables to avoid strong
tonal context effects, as well as intonation downshift on the last syllable of breath groups. Hooked wire EMG electrodes were inserted at the CT, Vocalis, and SH. Correct insertion was checked with various non-speech manoeuvres before and after the experiment, and periodically during this course. Subjects pronounced the 16 sentences (4 segmentals x 4 tones) at a normal speech rate, in 10 successive blocks. In order to correct insertion of the electrodes in the CT and SH could be achieved for 2 subjects, both male native speakers of Standard Chinese, born and raised in Beijing, aged 26 and 38, with no known speech pathology. Similar data were obtained for both. We use here the data from the first subject.

3. EMG PATTERNS
For each sentence, all repetitions were lined-up and time-normalized, using 2 reference events. This technique allow for averaging utterances on a wide domain, and for coping with speech rate variations. One utterance per sentence, the closest to the mean with respect to the duration between line-up events is used as reference. Patterns of CT/SH activity related to tone production are found to be stable across segmental variations. The time relationships of the patterns are found to be stable and consistent with respect to the rime —not to the entire word part of the syllable. This confirms that the rime is the domain of tone [6]. Patterns can be described as follows:
- tone 1: CT activity begins to increase at about 200 ms before rime onset, reaches a peak of moderate intensity at 75-80 ms before rime onset, and finally decreases to a steady level that is maintained until the end of the rime.
- tone 2: SH activity reaches a peak value 70-80 ms before rime onset. CT activity starts much later in the syllable than for tone 1, and is more continuous. It parallels the Fo contour, but precedes it by 75-80 ms.
- tone 3: SH activity is extremely intense for this tone. It begins to increase at about 100 ms before rime onset, and drops down a little before rime offset. There is no CT activity for tone 3 (the CT activity at the end of a target syllable must be related to the next syllable /zi/, in tone 4).
- tone 4: CT activity is very intense and parallels the Fo contour with a lead of 70-80 ms. CT peak activity occurs at about 45 ms before rime onset. A moderate concentration of SH activity consistently appears, centered a little before the mid point of the rime.

Note that what one may call "secondary activities" of the SH in tones 2 and 4 are found for both subjects. In order to show that these activities are tone-related, we have compared tone 2 or 4 to tone 1, where the smallest SH activity, presumably related, is observed. Comparisons were made at each point of time between sets of utterances (see [5] for details). The region where tones 1 and 2 significantly differ with respect to SH activity is the region where "secondary" SH activity is found before rime onset. Similar results obtain for tone 4 versus tone 1: Fo fall in tone 4 is assisted by SH activity. Interestingly, these EMG patterns explain puzzling phonetic data in running speech: for each tone, the lower its tone contour onset, and the longer a tone 4 syllable, the lower its tone contour offset [7]. This can only be the result of an active Fo lowering device for tones 2 and 4. Indeed, SH activity is such a device.

Let us see now how much Fo modelling comes close to these data.

4. MODEL COMMAND PATTERNS
The model we propose here is adapted from Fujisaki’s model for Japanese [2]: we use impulse commands to produce the "phrase component", which is assumed to represent the overall intonation, and step commands to produce local variations of Fo in the syllable domain. For Japanese, step commands are paired to form "accent commands":
o one step command followed by one offset step command of opposite amplitude. For Chinese, we call such pairs of commands "monad commands". We use both "positive" and "negative" tone commands; positive ones have an onset step command of positive amplitude and rime raise Fo, whereas negative ones have the opposite pattern and lower Fo. Time constants and damping coefficients characterize the responding
system. They are kept constant within a given utterance. However, the system is allowed to respond differently to onset and offset command responses, and to positive and negative commands. Critical damping is assumed for phrase commands but not for tone commands. Amplitudes and time locations of the commands characterize the excitation to the responding system. Practically, for a given utterance, the input to the model comprises the actual Fo data, and the initial estimates of excitation and system parameters. The latter are then optimized to minimize the discrepancy between the response of the system and the actual Fo data. Indeed, the optimization process does not lead to a unique solution. However, qualitative patterns of commands for each tone have emerged from our previous studies [4]. We use them as initial estimates, in order to reduce the search space of the optimization process. They may be summarized as follows: one positive tone command for tone 1, and, likewise, one negative command for tone 3, roughly spanning the whole range; one main positive command followed by a weaker negative one for tone 4, and the opposite pattern for tone 2. These patterns are qualitatively similar to the observed EMG patterns.

5. COMPARISON
We examined further the analogy by applying the model to the speech material described earlier. For each sentence, we analyzed the utterance that had served for time scale reference in the processing of EMG data. Care was taken to standardize analysis conditions for all utterances. In particular, parameters for the optimization process were the same for all utterances, and initial estimates were similar across segments. Fig. 1 shows CT and SH activities, together with the commands obtained for the segments /mi4/. Similar results obtain for other segments. Tone commands and CT/SH activities related to target syllables are compared with respect to their amplitude and their timing relative to the target syllable range. Results are summarized in Table I.

For timing, there is a good agreement between CT/SH activities and tone commands. Positive tone commands parallel CT activity, while negative ones parallel SH activity. However, amplitudes are poorly correlated.

6. DISCUSSION
EMG activity reflects an internal force developed within a muscle, whereas commands just indicate target Fo values. Contraction of the CT for example, produces a motive force f0, which tends to lengthen the vocal folds. The linear system approximation entails that f0 counteracts mechanical resistances to motion: inertia, friction, and elasticities. As simple mathematics can show, in order to raise Fo from a rest level to a high level, as in tone 1, rapidly enough to keep pace with the speech flow, f0 must overshoot the target value corresponding to the high level static equilibrium. When this level is reached, f0 drops down to the target value, and eventually fades away when the high level is given up. Hence, the typical profile of CT activity in tone 1.

That similar timing are observed for EMG activity and commands indicates that target values of Fo are programmed as target values of muscle tensions. Amplitudes of commands and EMG activities may correlate where Fo adjustments are stabilized, as after the onset of tone 1. Elsewhere, EMG amplitudes reflect dynamic aspects of Fo control, while commands reflect static equilibrium, that is, target Fo values.

REFERENCES
ACOUSTIC CORRELATES OF STRESS SHIFT

Stefanie Shattuck-Hufnagel

Research Laboratory of Electronics, MIT, Cambridge, Mass.

ABSTRACT

The prosodic phenomenon of stress shift, in which a stronger prominence is perceived on the first syllable of a word like "Mississippi" than on its main stress syllable "sip", in stress-clash conflicts like "Mississippi mud", has been attributed to rhythmic stress clash; the close approximation of two rhythmically prominent syllables is relieved by the leftward shift of the first prominence to an earlier syllable. Acoustic measures suggest that intonational prominence can play a substantial role in this phenomenon.

1. INTRODUCTION

The prominence pattern known as 'stress shift' has received considerable attention in the last decade, as one of the cornerstones of metrical phonology. Speakers of a number of languages judge that, under certain circumstances, the strongest prominence in polysyllabic late-stress words like "Mississippi" occurs not on the main stress syllable "sip", but on the earlier syllable "Miss". The prosodic environments that induce this apparent stress shift, as in "Mississippi mud", have been described as 'stress clash'. That is, the close approximation of two strong prominences, on "sip" and "mud", is rhythmically unacceptably, and is avoided by shifting the first prominence leftward to an earlier strong syllable [2, 8, 9, 11, 13, 17, 18, 20, 21].

Some speakers demonstrate elegantly systematic intuitions about the environments in which this apparent shift will occur (e.g. in American English, the main-stress-initial word "legislator" will induce the shift in "Mississippi legislation", but the secondary-stress-initial word "Mississippi" will not.) For other speakers, the facts are less clear, and even the existence of the phenomenon may be in question. Experimental measures have not yet produced convincing evidence to support the claims of shift-producing speakers whose intuitions are, none the less, remarkably consistent [4, 7, 12].

An additional complication arises from the existence of intonation models in which a pitch marker occurs early in the utterance of a declarative sentence [3, 5, 6, 10, 14, 16, 19, 22]. To what extent might this marker, when it occurs on e.g. the first syllable of "Mississippi mud", contribute to the impression that a leftward shift of stress has occurred?

In this preliminary study, one of a series of ongoing experiments designed to disentangle these issues, we explore two acoustic measures which might be expected to reflect the perceived shift in prominence: duration and F0. We confine the investigation to spoken prose (in Abercrombie's sense, distinct from conversational speech [1]) in American English, and we omit for now any discussion of the potentially important characteristics of intensity and loudness. For a limited number of sentences, we address the following question: In utterances for which both metrical theory and perceptual evaluation indicates an apparent stress shift, is there any evidence that either the F0 or the duration of the shift-receiving syllable reflects the change?

2. METHOD

Speech materials consisted of single words spoken in the frame sentence "Say the X again" and their candidate stress-shift counterparts "Say the XY again". The three words investigated, Mississippi, Massachusetts and Maxine, begin with voiced nasal-vowel syllables which permit both F0 tracking from the preceding word, and reasonably reliable measurements of segment duration. The corresponding stress shift; candidate phrases were Mississippi legislature, Massachusetts Avenue and Maxine Jones. A seventh phrase was included which was predicted to undergo stress shift because of the lack of stress clash: Mississippi legislation.

The seven stimulus sentences were produced as part of a larger set of utterances by nine speakers, five male and four female. The utterances were recorded on cassette tape, in a partially sound-attenuated room, and digitized at 10,000 kHz. Duration measures were taken by hand from cursor readouts on waveform displays, and F0 estimates were obtained automatically by a procedure developed by Dennis Klatt that involves finding the spacing between the harmonics in the spectrum.

Perceptual evaluation of stress shift in the resulting 63 utterances was carried out by the author. In many cases the outcome was clear: either the largest prominence was on the first syllable of the target word, (i.e. stress shift had occurred), or it remained on the syllable which would normally carry main lexical stress (i.e. no stress shift had occurred.). Interestingly, a third pattern emerged, in which the initial syllable and the main-stress syllables of the target word seemed to be of equal prominence. These cases were labelled 'unclear', and were analysed separately.

3. RESULTS & CONCLUSION

Perceptual analysis: Of the 27 target words predicted to undergo shift, 14 were judged to be shifted, while 2 had their major prominence on the mainstress syllable and thus had not shifted; both of the latter were utterances of 'Say Maxine Jones again'. In the remaining 11 cases the relative prominence of the first and main-stress syllables of the target word was judged unclear.

Of the utterances of 'Mississippi legislation', predicted not to undergo shift, 6 were shifted and 3 were unclear. Finally, of the utterances of the single target words Mississippi, Massachusetts and Maxine in the frame sentence, 25 were unshifted and two were unclear.

Thus, single words did not undergo shift, just over half of the candidate shift words did, and the phrase "Mississippi legislation", predicted not to shift, was perceived to shift more than half the time.

Individual speakers were somewhat consistent: five speakers shifted 4 or 3 utterances, and four shifted 1 or none. Individual sentences were also somewhat consistent, shifting for 5, 6, 5 and 4 of nine speakers. This pattern of results suggests the wisdom of perceptually evaluating candidate shift utterances to determine whether or not stress shift has occurred, before analysing its acoustic correlates.

Duration analysis: For each speaker, the duration of the first syllable of a target word produced alone in the frame sentence was compared with its duration in the shift candidate context, and the results tabulated separately for shifted, unclear and unshifted utterances. No striking differences among the 3 distributions were noted (Fig. 1a), perhaps because of variation in speaking rate from utterance to utterance. If stress shift is accompanied by systematic shifting of the stressed syllable, the differences (as other investigators have reported) are not easy to demonstrate with a simple comparison between utterances.

F0 analysis: The F0 results present a somewhat clearer picture. We report here only the within-word measures of F0 change in the first syllable of the target words. This was defined as the distance and direction of the change between the highest and lowest F0 values in the syllable. In words judged to show stress shift, the change was generally large and positive, ranging up to 71 Hz, while the unclear cases were more often small or negative. Finally, the 2 cases judged to be unshifted, with their major prominence remaining on the mainstress syllable, showed large negative changes in F0 in the first syllable: -36 and -15 Hz. The distribution of F0 changes in the initial syllable of the target words is summarized in Figure 1b.

These results suggest that utterances in which stress shift is perceived tend to have large F0 rises in the shifted-to syllable, although such a rise is apparently not sufficient to ensure the perception of stress in all cases, since a subset of
or
This inference markers pattern duration increases: the examples corresponding by prominent, cases "Mississippi" syllables where related to the perception of stress from the mainstress syllable fall the stress-shifted utterances the word, or the stress-shifted utterances of these Speaker, labelled this speaker shows that no possible interpretation of this this speaker relies on duration increases: the initial syllable of the target word was 30-50 ms shorter the stressed-utterances than in the corresponding single-word utterances. Other possibilities include a change in F0 from the last syllable of the preceding word, a relative F0 change (or relative duration) of syllable 1 and the mainstress syllable. The single-word cases for this speaker show a substantial fall in the first syllable of the target word (30-50 Hz), so that the stress shift cases always have a lesser fall in F0 in the shift than the single word cases, but it is unclear whether this fact is related to the perception of stress shift.

An interesting aspect of the initial-syllable patterns is the pitch marker observed in initial-syllable words in frame sentences. An example is shown to the left in Figure 2, where the initial syllable "Mi-" shown F0 rise for both "Mississippi" and "Mississippi legislature." Since no stress shift was perceived in the single-word case for this speaker, the initial-syllable marker is apparently overridden by a more prominent marker on the mainstress syllable "sip." This inference is supported by the F0 pattern for the mainstress syllable in the same word, shown to the right in the figure. A possible interpretation of this pattern is that speakers have two separate options for the placement of pitch markers on a polysyllabic target word: they can mark the initial syllable or not, and they can mark the mainstress syllable or not. On this view, the combination of pitch marking on the first syllable and no pitch marking on the mainstress syllable could contribute substantially to the perception of stress shift. For a synthesis algorithm compatible with this hypothesis see references [15].

**Conclusion:** The preliminary results reported here illustrate several significant points: (1) it is important to evaluate stress shift cases for utterances perceptually before measuring possible correlates of stress shift, since not all claus contexts invariably induce shift and it occurs in some non-clausal contexts; (2) in some shift cases, the greater perceptual prominence of the shifted-to syllable may be a matter of intonational rather than rhythmic prominence; the hypothesis that this prominence early in the word reflects in part an unmasking of the prominence associated with an onset intonational marker on an earlier syllable of the word, an unmasking which results from the disappearance of phrasal prominence from the mainstress syllable (in favor of a later word), requires further testing, and (4) speakers can take different approaches to the problem addressed by stress shift models; for instance, speakers will take the options available to speakers to understand the relation between not only rhythmic and intonational aspects of prosody, but also lexical and prosodic prominence.

**Figure work:** Clearly, an understanding of stress shift will require a comprehensive approach involving phonological, acoustical, and perceptual analyses, with more speakers, more utterances, and more listeners doing the perceptual evaluations [4]. In addition, an important control experiment remains to be run. If the longer string of syllables in the stress shift candidate sentences causes the speaker to reach a higher early F0, the results reported above would have a very different interpretation. A control experiment comparing F0 and duration changes for initial syllables in non-stressful pairs like "manageable" vs. "manageable legislators" will test this possibility.

**4. ACKNOWLEDGEMENTS**

This work was supported by grants NSF IR1-8805680 and NIH 8-301-DC00075. Conversations with P.J. Price and M. Ostendorf have been invaluable.
ABSTRACT

Analysis of six regional Danish varieties, two Swedish and two German ones reveals striking differences in cues to the terminal or non-terminal function of the utterance, differences which are coupled with the absence or presence of separate signals to utterance completion.

1. INTRODUCTION

- This is a presentation of a part only of a comprehensive study involving also sentence accents, stress group patterns and final lengthening. It is further restricted to a display of only five (exemplary) varieties of the ten investigated. Even so, the presentation will of necessity reduce to a summary of the results and the ensuing discussion. A full account can be found in <1>. To save space, the figures are highly compressed.

2. RESULTS

2.1 Global versus local

- The criteria for categorizing signals to terminal and non-terminal intonation, respectively, as local versus global, are as follows:
  - Local cues: (1) the last stress group does not deviate in any principled way from preceding ones. (2) It forms the termination of one smooth overall course whose slope varies with utterance length (less steep in longer utterances, ceteris paribus) and with terminal vs. non-terminal intonation (less steep when non-terminal, ceteris paribus). See (b).
  - Local and global signals may co-exist if final cues are preceded by significant global differences.

- Varieties with global cues to terminality (Copenhagen, Nestved, Aalborg, Tønder (Danish) and Kalmar (Swedish)) do not have default accents, and signal focus by stress reduction rather than by sentence accents proper. It is also void to postulate any 'final lowering' gesture for Copenhagen (and the other 'globales').

- The local varieties are Bornholm, Sønderborg, Flensburg, North German, and Stockholm. (Sønderborg is not exposed here.) I have counted North German among the local types, but it seems in fact to constitute a hybrid between global and local: prelude slopes in long vs. short terminals and in terminals versus non-terminals do differ.

2.2 Is terminality coincident with completion?

(a) Utterances with final (or no) accent

- Fallbacks in Stockholm are uncontroversially separate completion signals, tagged onto the sentence accent rise. The terminal vs. non-terminal cue lies in the preceding accented syllable, which is higher in non-terminals vs. (c) broken vs. solid line. Lowering finally seems to be the only option for completion in Stockholm.

- In Bornholm, terminal and non-terminal contours are different only by the movement of the last post-tonic in the final stress group, cf. (d, e). Thus, final falls signal terminality as well as completion, and final rises likewise simultaneously signal both non-terminality and completion. However, final falls and rises reach the same low or high offset value, irrespective of their onset level (which is a matter of accentuation), which indicates that a separate completion command is involved.

- The two German non-terminals, (g) and (h), share an overall slope which is less steep than in terminal utterances of comparable length and (g) further more has a final post-tonic rise, whereas in (f) the utterance ends with a 'low'.

- German non-terminals, when the latter are succeeded by a completion 'low' provide us with a somewhat counter-intuitive situation where non-terminals have larger final falls than terminals. This ambiguity is resolved when (1) the final low, and thus the descent is assigned to utterance completion and (2) the level of the last stressed syllable, which determines the magnitude of the fall, cues terminality. The level of this last stressed syllable follows from differences in global slopes.

(b) Utterances with non-final accent

- If the highs and lows described above are indeed separate completion signals they must stay in place, at the end of the utterance, even if sentence accents and terminality cues move back. They should then be reached either progressively or via a discontinuity in the preceding Fo course.

- Stockholm has only low completion cues and maintains an unmistakable low in final position: The post-accentual course can be regarded as a smooth continuation between the early accent peak and the utterance fall, with diminished word accents superposed, cf. (b).

- In Bornholm, the final point in terminals constitutes the end of a generally smooth fall from the early accent, cf. (l). The fall from the high accented syllable in the non-terminal is not as damped as further movement is suspended until the final rise, cf. (j).

- In German, like in Bornholm, the initial accent is
succeeded by a fall, which must be considered part of the accent command. In non-terminals, further movement is suspended, until the very final gesture, which may be either rising or falling, to the completion high or low, respectively, cf. (k, l). In terminals, the fall is continuous through the post-accented syllables until the slight skip up at the end to punctuate the final low, cf. (f: broken line). The same situation thus holds as for final accents, apparently. I.e., non-terminals may be doubly cued, partly by the higher course of the post-accentual tail, partly by the final completion rise, or merely by a higher post-accentual stretch, which magnifies the final fall to the completion low.

3. CONCLUSION
- Insofar as the acoustic cues to terminal or non-terminal and to utterance completion may be separate in time (located in different places in the utterance) they must have separate representations in the prosodic system. This existence of two separate commands is supposedly maintained if and when terminality and completion pile up in the same location, as they do in utterances with final (or no) sentence accent.
- Separation of terminality and completion is unambiguous in Stockholm. The completion is always low, and the cue to terminality is always associated with the sentence accent rise, independent of its location.
- In Bornholm terminality is bipartite. There is a cue at the very end, in the movement of the last syllable, the completion cue. But there is also a difference in the magnitude of the fall from an early accent, which is deeper in terminals than in non-terminals.
- German operates in a similar fashion to Bornholm except for the interesting fact that non-terminal and terminal is not inextricably connected with high vs. low completion: The low completion does not unambiguously also cue terminality.

<1> Gronnum, N. (in print), "Prosodic parameters in a variety of Danish Standard languages, with a view towards Swedish and German", Phonetics.
INTONATION MODELLING IN A TEXT GENERATION PROGRAM
Carlos Gussenhoven, Toni Rietveld and Lex Elich
Department of English & Department of Language and Speech
University of Nijmegen/The Netherlands

ABSTRACT
In this contribution we describe the implementation of an autosegmental description of intonation in Dutch.

1. INTRODUCTION
We present a brief description of the implementation of an autosegmental description of the intonation of Dutch in an allophone synthesis system. Our longer-term aim is for this program to be part of a text-generation program. At present, we assume the following as given:

1. A phoneme string, with word and syllable boundaries; the latter are symbolized by '.
2. Accents, symbolized * before accentuated syllables
3. Association Domain (AD) boundaries, symbolized (...);
4. Scaling Domain (SD) boundaries, symbolized (...)
5. A phonological transcription of the intonation contour.

In (1) we give the representation of *Jan (girl's name), without the information under 5.

(1) (f m A r - * A n )

Our program translates the phonological transcription into a string of targets. A target is a point defined by (a) an F0-value (Hz), and (b) a time value relative to the phoneme string. The intonation contour is obtained by interpolating between all targets. We will first describe the phonological representation of the contours, then given the timing rules, and finally the F0-rules.

2. THE PHONOLOGICAL REPRESENTATION
The representation is built up by inserting intonational morphemes (see also Gussenhoven [2,3]).

Pitch Accents. Minimally, each * in the string must be provided with one of three PITCH ACCENTS: H0, LH, or HH. The choice is free, except that all nonfinal *'s in an AD must have the same Pitch Accent.

In addition, a larger number of optional morphemes are available. These specify AD's, or, in one case, the AD:

- Accental Downstep. An AD may be provided with ACCENTUAL DOWNSTEP. This means that the H of each non-initial FI will be realized with lower pitch than the preceding H. ACCENTUAL DOWNSTEP is symbolized by placing the diacritic ] before the AD.

- Accental Downstep with Spreading. An AD which has been provided with DOWNSTEP, may additionally be provided with SPREAD. This means that each non-final FI will be realized as a plateau instead of a peak, while the final FI is rewritten L, and the preceding L is deleted. SPREADING-DOWNSTEP is indicated by means of -1 before the AD.

- Narration. NARRATION is applicable to FL and LH, and causes their T to spread. Thus, while an unnotated realization of FL will show a peak at the accented syllable, a narrated one will show a high plateau beginning at the accented syllable and ending just before the next *, or the end of the AD. Narration cannot occur in an AD with downstep. It is symbolized by placing ] before the AD.

Modifications. Pitch Accents come in a number of variant forms, which are described as 'modifications' of the basic shapes. Two modifications are implemented. The first, DELAY, causes the association of the Pitch Accent to be shifted rightward. DELAY is implemented as the prefixation of a L tone segment, which is timed like the T of the underlying Pitch Accent; it is symbolized by placing ] before the AD. The second modification, HALFCOMPLETION, is possible only for the last Pitch Accent of an AD. It causes the contour to end on mid pitch. It does not combine with either NARRATION or DOWNSTEP. It is symbolized by placing - before the AD.

3. IMPLEMENTATION: Timing of Targets
The domain for the timing rules is the AD. That is, the first and the last frame of the AD act like firm walls, and rules cannot locate targets beyond them. Accented syllables provide AD-internal reference points. The notation T is used for tone segments other than %T, T, or T'. Each tone segment yields either one or two targets. The %T, the last T of an AD-final Pitch Accent, and a spread F translate into two targets. All other tone segments translate into single targets. Where a tone segment yields two targets, the earlier one is called Target1 and the later one Target2. For ease of exposition, we here distinguish three levels of representation. First, there is the segmental string, with *'s added before the accented syllables, and (...) and (...) in place. Second, the temporal string, consisting of a series of frame numbers associated with the successive segments. (We here assume for the sake of convenience that a frame represents 10 ms.) The third representation is the tonal string.

We show the timing of targets with the help of association lines between tone segments and the frames. Our rules contain a number of timing parameters, whose values can be manipulated in order to explore the perceptual effects of different timings of tone segments. We give default values which have informally been observed to give reasonable results.

- ∪: locate Target at a distance of STARTTIME of the vowel duration, counting from the beginning of the vowel. (Default STARTTIME = 50%)

- ∪: a spreading ∪ (for SPREAD or NARRATION) has two targets: Target as above
Locate Target2 at TOTIME from frame associated with next \( T \). (Default TOTIME=100ms.) If AD is narrated, locate Target2 at (SPREADTIME)*TOTIME from next * or . (Default SPREADTIME = 1.3).

- \( \%T \): associate Target1 with first frame of AD; Locate Target2 at TOTIME before *. (Default TOTIME=100ms)

For non-final Pitch Accents:
Each \( T \) is located at FROMTIME from the distance between preceding target and * it is less than FROMTIME + TOTIME, locate Target midway between preceding target and *. The last \( T \) is located at FROMTIME from next *, if the distance between preceding target and following * is less than FROMTIME + TOTIME, position Target midway between preceding target and *. (Default FROMTIME=100ms)

For final Pitch Accents:
All T's except the last as above. The last \( T \) receives two targets:
Locate Target1 at FROMTIME after *. Locate Target2 at TOTIME before the end of AD, if \( T \%F \) follows. If no \( T \%F \) follows, associate Target2 with last frame.

- \( \%T \%F \): associate Target with last frame

Where the space provided by the segmental string is less than FROMTIME + TOTIME, Target2 may inappropriately be timed earlier than Target1. In such a case, OVERLAP and SIMPLIFY apply so as to associate Target1 with the frame that lies midway between them, and to delete Target2.

In (6), we illustrate a situation in which the available time is less than FROMTIME and TOTIME. Representation (6a) results from applying the first timing rule (STARTTIME). In (6b), we see the result of the other timing rules without OVERLAP. Target2 of \( \%T \%F \) was positioned by going TOTIME leftward from *, and hitting the lefthand boundary of AD (cf the dotted association line). It is thus associated with the same frame as Target1. For Target1 of \( L \), we count FROMTIME from *. For Target2, we count TOTIME back from the right-hand boundary. Notice that the two targets overlap, as shown in the added target tier. Their associations are given as dotted lines to indicate their provisional status. Representations of the state of affairs after the application of OVERLAP and SIMPLIFY. This representation is ready to go to the F0-rules.

### 4. IMPLEMENTATION: F0

The calculation of F0-values is performed by an implementation model described in Van den Berg et al. [1]. This model is a modified version of that proposed in Ladd [5]. Briefly, it provides a high reference value (which equals that of the first \( T \)) and a low reference value (which equals that of \( L \)), together defining a register, whose width is referred to as TRANGE (i.e. the distance between \( T \) and \( L \)). The starting values are determined by three parameters that are intended to model speaker-to-speaker variation in general pitch height, and different degrees of prominence and liveliness. Their settings remain in force throughout the SD. An Accenntual Downstep factor \( da \) determines the lowering of \( T \) targets in an AD with accentual downstep. The distance between \( L \) and the most recent F0-value for a (downstepped or undownstepped) \( T \)-target is referred to as TRANGE. For targets after undownstepped \( T \), TRANGE equals TRANGE.

A Phrasal Downstep factor \( dp \) determines the lowering of AD's in an SD with phrasal downstep.

For targets other than those of \( T \), we can be flexible in the sense that not only the high and low reference values will be used, but any intermediate value. That is, we refer to values around the reference values by means of percentages, in the manner of Home [4].

#### 4.2. F0-rules

- \( \%T \%F \): Target1: \( L = \) STARTSINK of TRANGE. (Default STARTSINK=35%);
- Target2 = (STARTSLOPE)*Target1 (Default STARTSLOPE = .9)
- \( T \%F \) (high reference)
- \( L \%F \) (low reference)
- \( \%F \%T \) F0 as given by the Accenntual Downstep factor \( da \).
- \( L \%F \) in final Pitch Accent = \( L \%F \) (Target1 end)
- \( T \%F \) if HALF-COMPLETION is in effect, delete Target1 and scale Target2 at HALF of TRANGE (Default HALF=60%).
- L in non-final Pitch Accent = SAG of TRANGE (Default SAG = 23%)
- \( \%L \%F \) = ENDSINK of TRANGE (Default ENDSINK = -10%)
- \( IP \) = previous Target + (ENDRISE of TRANGE) (Default ENDRISE = 30%).

#### 4.3. The F0(m,n)-module

The implementation model F0(m,n) is given bellow. It calculates the F0-value for the nth \( T \) in the nth AD.

\[
F0(m,n) = \frac{1}{\text{Nw} \times \text{Wn} \times \text{dp} \times \text{Sp} \times \text{m-1} \times \text{n-1}}
\]

**Parameters:**

- \( \text{Sp} = 4 \) if Phrasal Downstep, 0 if not;
- \( \text{Sp} = 4 \) if Accenntual Downstep, 0 if not;
- \( \text{T} = 1 \) for \( L \%F \), and -1 for \( T \%
- \( \text{Fr} \) = Reference line at the bottom of the speaker's range (default: 50 Hz for men and 100 Hz for women)

\( \text{N} \) = Defines the range, or the mean starting value above Fr (Default: 2)
\( \text{W} \right) \) = Determines the distance between \( L \%F \) and \( L \%F \) (Default: 1.6)
\( \text{dp} \) = Downstep factor for downstepping \( T \) targets within the AD. ("Accenntual Downstep". Default: 0 if \( S = 1 \), and .70 if \( S = 0 \))

**REFERENCES**


WAYS OF EXPLORING SPEAKER CHARACTERISTICS AND SPEAKING STYLES

Björn Granström and Lennart Nord

Dept of Speech Communication & Music Acoustics, Royal Institute of Technology, KTH, Box 70014, S-10044 Stockholm, Sweden.

Phone 46 8 7907847, Fax 46 8 7907854

*names in alphabetic order

ABSTRACT

In the exploration of speaking style and speaker variability we make use of a multi-speaker database and of a speech production model. A recent version of this model includes a variable voice source and a more complex modelling of the vocal tract. Systematic variation in speech synthesis has been used as a tool to explore possible style and speaker dimensions. Preliminary listening experiments have been carried out with the aim to investigate whether it is possible to describe different synthesis samples according to different attitudinal and emotional dimensions.

1. INTRODUCTION

An increasing amount of knowledge concerning the detailed acoustic specification of speaking styles and of speaker variability is presently accumulating. The ultimate test of our descriptions is our ability to successfully synthesize such voices [1]. A better understanding will also have an impact on several applications in speech technology. A systematic account of speech variability helps in creating speaker adaptable speech understanding systems and more flexible synthesis schemes.

Why introduce emotional content in speech synthesis? Firstly, to increase naturalness and intelligibility of a spoken text. Speaking style variation and the speaker's attitude to the spoken message are also important aspects to include. However, if the attitude can not be explicitly signalled, it is better to stick to a more neutral, even non-personal, machine-like synthesis.

Several applications can be foreseen, e.g. synthesis as a speaking prosthesis where the user is able to adjust speaker characteristics and emotional content or in translating telephony, where speaker identity ought to be preserved and tone of voice aspects also form part of the communication.

2. NEW TOOLS

In the exploration of speaking style and speaker variability we make use of a multi-speaker database and speech database project we have started to collect material from a variety of speakers, including professional as well as untrained speakers [2]. The structure of the database makes it possible to extract relevant information by simple search procedures. It thereby becomes easy to retrieve information on the acoustic realisation of a linguistic unit in a specified context. Earlier studies have concentrated on linguistic structures rather than paralinguistic descriptions. We aim at explicit descriptions that are possible to test in the framework of a text-to-speech system [3]. A recent version of the speech production model of the synthesis system includes a variable voice source and a more complex modelling of the vocal tract [4]. This synthesis model gives us new possibilities to model both different speakers and speaking styles in finer detail. The necessary background knowledge, however, is in many respects rudimentary. We will here show one example of analysis-synthesis applied on emotive speech.

3. ACOUSTICS OF EMOTIONS

In acoustic phonetic research most studies deal with function and realization of linguistic elements. With a few exceptions, e.g. [7,8], the acoustics of emotions have not been extensively studied. Rather, studies have dealt with the task of identifying extralinguistic dimensions qualitatively and sometimes also quantitatively by using scale methods. Spontaneous speech has been used as well as read speech with simulated emotional expressions. Judgements have been made by the researchers' ear and also by a variety of listening tests, using untrained groups of listeners. An interesting alternative is to ask the listener to adjust perceived stimuli to some internal reference, such as joy, anger etc. This is typically done by using synthetic speech, which cannot be too poor in quality if emotions should be conveyed. Recent experiments using DECTalk has been reported by Cahn [2]. The amount of interaction between the emotive speech and the linguistic content of a sentence is difficult to ascertain, but has to be taken into account. It is not easy to define a speech corpus that is neutral in the sense that any emotion could be used on the sentences. Also some sex related differences might be observed. A study by Oster & Risberg [6], female joy and fear were more easily confused than for male voices, where instead joy and anger were more often confused by young listener groups. Also concepts like joy, anger etc. can be expressed very differently and a unique perceptual - acoustic mapping is probably impossible.

Note that the voice does not always give away the complete speaker attitude. It is often observed that misinterpretation of emotional cues occurs if the listener is perceiving the speech signal without reference to visual cues. Depending on the contextual references it is thus easy to confuse anger with joy, fright with sorrow, etc.

4. SPEECH ANALYSIS

We have analysed readings by two actors who were portraying different emotions by reading a fixed set of sentences in different ways: with anger, joy, fear, sadness, surprise and also in a neutral tone of voice. This material has already been used by Oster in the investigation referred to above [6], with the aim of investigating the possible differences in

ability to perceive emotion acoustically, as shown by two listener groups, young hard-of-hearing subjects and young normal hearing subjects. We specifically analysed pitch, duration and segmental qualities and also made rhyme matching of these sentences trying to extract the relative importance of the different acoustic cues.

One example from the database can be seen in Figure 1, where two versions of the Swedish sentence 'De kommer på torsdag' (They will arrive on Thursday) pronounced by a male actor in an angry and a joyful mode are shown. Numerous differences can be observed. For this particular 'angry' utterance the pitch is lower and more even than the "happy" utterance. The voicing is also stronger and somewhat irregular especially in the first vowel (probably the false vocal cords are also involved).

For some of the sentences it was obvious that the two actors made use of a number of extra acoustic prototypical breaks and jitter, lip smacks, etc which often contributed in a decisive way to the intended emotion. This means that a standard acoustic-mimicry analysis of sentences with different emotional content, in terms of e.g. duration, intensity and pitch, does not discriminate between emotions, if the speaker relies heavily on non-phonetic cues in the production.

As a point of reference we have also initiated a small study on spontaneous speech from radio interviews. This speech often contains passages that are extremely compressed or expanded. These effects are difficult to make use of in speech synthesis and in natural speech communication. Nonetheless, it is a good reminder of just how diverse and flexible the speech signal appears in real-life communication.

5. VALIDATION BY SYNTHESIS

Different analysis-by-synthesis techniques show great promise in deriving data for the synthesis of different voices, speaking styles and emotions. Specifically, we investigated an interactive production paradigm. We asked subjects to sit at a computer terminal and to vary the horizontal (X) and vertical (Y) position of a point within a square on the screen by means of a mouse. The X and Y values can be used in a set of synthesis rules,
changing e.g. different aspects of the voice. In this way we set up a number of rules that changed e.g. pitch deviations, intensity dynamics or voice source parameters of a synthesized sentence. The subjects were asked to try different combinations of these parameters by moving the mouse and reporting on the impression that the synthetic sentence made on them in terms of e.g. emotional content. In Figure 2 a result from such an experiment is shown. The X dimension corresponds to the slope of the declination line where a low coordinate value, (left), corresponds to a rising contour and a high value, (right), corresponds to a falling contour, with the midpoint in the sentence kept as a constant pitch value. The Y dimension is the pitch dynamics, where the low end corresponds to small pitch movements and the top to larger local pitch excursions. The tested sentence is the same as in Figure 1, i.e. a linguistically quite neutral statement. Obviously, the variations suggest several different attitudes to our listeners. The task appeared quite manageable to the subjects, who responded with a fair degree of consistency. We are pursuing this line of experiments further including also voice source variations.

**REFERENCES**


ANALYSE DE LA PROSODIE DE LA PAROLE SPONTANÉE EN SUÉDOIS ET EN FRANÇAIS

P. Touati
Institut de Linguistique et de Phonétique, Lund, Suède.

ABSTRACT

This paper reports on a methodology developed to study prosody in spontaneous speech, incorporating four different kinds of analysis: (1) analysis of the discourse structure of the speech corpus without specific reference to prosodic information, (2) auditory analysis in the form of a prosody-oriented transcription, (3) acoustic-phonetic analysis and (4) analysis-by-synthesis. Analysis (3) is illustrated with examples in spontaneous Swedish and French.

1. INTRODUCTION

Cette communication présente une méthodologie élaborée au cours d'un projet de recherche consacré à la prosodie de la parole spontanée en suédois, grec et français (cf. en particulier [2] et [3]). Notre effort inaugural a été d'intégrer dans une même démarche expérimentale des sources de connaissances diverses susceptibles de permettre l'analyse d'un corpus de parole spontanée dans un espace qui s'étend de la description discursive des catégories prosodiques à une transcription de la parole au niveau de la parole phonétique. Notre méthodologie a été réalisée disposable à la prosodie, permettant de faire émerger certaines contraintes discursives et de faire une évaluation de la prosodie à travers l'interprétation des locuteurs et de la réalisation de tours de parole. La description de la structure discursive est énumérée ainsi que les différents textes, leur articulation successive, les rapports de domination entre les locuteurs au fil des répliques et la réalisation de tours de parole en termes de prise de parole, de passage de parole... Cette description est ultérieurement mise en relation avec l'organisation prosodique.

2. METHODOLOGIE

2.1. Analyse discursive

Effectuée sans référence particulière à l'organisation prosodique, cette analyse a pour fonction de faire émerger certaines contraintes discursives tels que l'organisation temporelle, l'émergence de locuteurs et la réalisation de tours de parole. La description de la structure discursive énumère ainsi les différents textes, la description de la lecture et dans des conditions de communication authentiques.

2.2. Analyse auditive

L'analyse auditive précède à un décodage prosodique du corpus. Elle se traduit tout d'abord par une transcription orthographique dont les hésitations, les rires, les chevauchements de tours de parole etc. (cette transcription est préalable à l'analyse discursive). La transcription prosodique a essentiellement pour but de mettre en évidence la manière dont les fonctions démarcatives et hiérarchiques ont joué dans la structuration du corpus. Cette transcription est donc

selective. Les cinq catégories sélectionnées participent, quoique de manière différente, à la réalisation de ces fonctions. Ces catégories sont la proéminence accentuelle, le regroupement prosodique, le registre de voix, la marque des frontières de mots et les pauses. La transcription est également abstraite dans la mesure où ces catégories sont loin d'avoir la même manifestation acoustique dans chaque langue et où une même catégorie est manifestée par plusieurs paramètres acoustiques. Le choix des symboles de transcription suit si possible les recommandations d'IPA, autrement il s'aligne sur un critère de transparence iconographique (cf. Tableau 1).

2.3. Analyse acoustico-phonétique

Cette analyse précède au décodage acoustico-phonétique du corpus (pour les critères concernant le choix des corpus et la procédure expérimentale cf. [2], [4] et [7]). Des cinq catégories auditive, seules les pauses silencieuses relèvent clairement de la dimension temporelle du signal. La catégorie 'regroupement prosodique' délimite les domaines d'exercice des trois autres catégories qui sont liées aux variations verticales de Fo. La modélisation des tracés de Fo permet une première analyse qualitative des données obtenues. Elle s'opère en assignant aux valeurs-cibles minimale et maximale de Fo des représentations phonologiques intermédiaires en termes de segments tonaux Hi(high) ('Haut') et Lo(low) ('Bas'). Les représentations phonologiques intermédiaires des accents du suédois et du français ainsi que leurs points de synchronisation syllabiques représentés par les symboles de transcription sont exemplifiés ci-dessous en (1) et (2). Dans les deux langues, le segment tonal synchronisé avec la voyelle accentuée est décoré d'une étoile. Ces représentations sont également intégrées dans les tracés présentés dans l'annexe.

(1) Suédois

accent non-focal

accent I

accent II

accent focal

(2) Français

accent non-focal

accent I

accent II

accent focal

2.4. Analyse par synthèse

L'analyse par synthèse effectuée jusqu'ici présente eu pour objectif d'évaluer et de caractériser la valeur textuelle et interactionnelle de certaines configurations tonales (cf [4] et [7]).

3. EXEMPLES

3.1. Lecture spontanée en suédois

En suédois, le rôle de pivot joué par l'accent focal — il détermine l'absence ou la présence d'une séquence de tons abaissés ("downstepping") — a été mis en évidence dans la lecture (cf.[1] et Fig.1). Les accents situés en position post-focale se caractérisent par un abaissement tonal successif. En revanche, les accents situés en position pré-focale ne montrent aucun abaissement tonal; ils se caractérisent par une proéminence tonale plus ou moins égale. Il est intéressant de noter que les données du spontané constituent ce rôle de pivot joué par l'accent focal. Un exemple d'abaissement après un accent focal initial est présenté à la figure 2:1 et un exemple de non-abaissement avant un accent focal final à la figure 2:2.
3.2. Accentuation chez un enfant français

L'échantillon de corpus étudié a montré la manière dont l'accentuation joue dans la structuration prosodique interne au tour de parole chez l'enfant. En règle générale, c'est une montée tonale LH* qui est associée aux syllabes accentuées des groupes en position non finale de tour de parole (cf. Fig. 3:1 ("pas"), Fig. 3:2 ("vie dans mourir") et Fig. 3:3 ("vieux"). Cette montée tonale est combinée de manière relativement stéréotypée avec une pause interne. Les groupes situés en position finale de tour de parole se caractérisent par une descente tonale graduelle D et un segment tonal L* sous la dernière syllabe accentuée (cf. Fig. 3:3 "et ben on 'meurt'"). On constate peu d'occurrences d'accents focal.

3.3. Registre de voix chez un politicien français

Le corpus étudié a mis en évidence l'importance du changement de registre de voix dans les débats politiques dans les masses média. La spécificité de ce genre de communication poussent les participants à produire de la monologie (en haut) et des débats politiques dans le registre de long mélodie. Les politiciens étudiés sont répartis en trois catégories, selon leur manière de produire de prosodies (en haut), configuration tonale (au centre) et transcription orthographique avec marqueurs prosodiques (en bas); Voir § 3.3. pour des explications concernant les symboles (HL) utilisés dans la figure.

4. REFERENCES


5. ANNEXE

Figure 1. Abaissement (à gauche) et non-abaissement (à droite) tonal dans un dialogue spontané en suédois; effet d'un accent focal initial et final; onde sonore (en haut), configuration tonale (au centre) et transcription orthographique avec marqueurs prosodiques (en bas); Voir § 2.2. et 2.3. pour des explications concernant les symboles (HL) utilisés dans la figure.
INTONATION PATTERNS IN GREEK DISCOURSE

Antonis Botinis

Department of Linguistics, Athens, Greece
and Dept. of Linguistics and Phonetics, Lund, Sweden

ABSTRACT

The object of this investigation is a classification of discourse intonation patterns in spontaneous Greek discourse. Our goal is to describe the distribution, manifestation, and function of discourse triggered accents such as ‘initiative’, ‘completive’, and ‘concessive’. Furthermore, pitch range as a discourse turn and topic regularity along with inter-speaker pitch adjustment is touched upon and the notion of ‘pitch-concord’ is introduced.

1. INTRODUCTION

This presentation is about intonation patterns in spontaneous Greek dialogues, originally outlined within the framework of the ‘KIPROS’ project, which is the Swedish acronym of a research program on contrasive and interactive prosody [5]. We started our investigations by first analysing a face-to-face spontaneous conversation between two speakers, with emphasis on certain local conditions in the laboratory. The next step was an investigation of regular Fo- and contour patterns which make up the skeleton of discourse prosody [2], by examining the intonation structure of two speakers involved in a telephone conversation where somatic communicative means are excluded in favor of prosodic ones. Our current research program is on the one hand the description of intonation patterns in a telephone context with more than two speakers at a time and, on the other hand, the global organization of intonation in spontaneous spoken discourse [4]. In the present paper, our main emphasis is on a selection of highly recurrent local patterns and their communicative function.

2. EXPERIMENTAL DESIGN

2.1. Speech Material

The present data consists of four short telephone conversations recorded from a local Athenian radio station. They are conversation extracts from an entertainment program, in which listeners may phone in and participate in a contest with the chance of winning various small presents. The topics of the conversations are related to the degree of the program participants’ successful answers and may be organized in subtopics; occasionally, topics outside the question—answer paradigm may appear in the course of a conversation. The speakers of our selection include two relatively young program leaders, a male and a female, and four program participants: two adults and two children, a male and a female respectively.

2.2. Speech Analysis

Our methodology includes four kinds of analysis: (1) analysis of the discourse structure in terms of topic development and turn-unit interplay; (2) auditory analysis and prosodic transcription (3) acoustic-prosodic analysis and (3) analysis-by-synthesis (see [2] and [6]). Here, we shall confine ourselves to the acoustic-prosodic analysis.

The acoustic-prosodic analysis consists of observations and classification of intonation patterns, and, of particular interest, stereotyped intra and inter-speaker pitch sequences characteristic of spoken discourse. This has been done through an interactive examination and listening of the relevant pitch contours on a VAX/VMS 11/730 computer system with the API program of the ILS package.

3. DISCOURSE PITCH PATTERNS

3.1. Specific Pitch Patterns

For the descriptions of intonation patterns in spontaneous speech, we will introduce a methodological distinction between stress and accent, prosodic terms which are often overlapping and/or interchanged in the current, prose-like literature. By stress we mean prominent syllables with no reference to pitch whereas by accent we do mean pitch gestures whether they are associated with stress or not. In the former case, stress alternations make up the rhythm of the language whereas, in the latter, the interconnection of accent is within the realm of intonation. Laboratory speech has taught us that stressed syllables are not necessarily assigned pitch gestures. This has been observed in previous material as well as in the present material. On the other hand, discourse oriented pitch gestures may be carried by unstressed syllables in specific environments with high communicative value, such as the beginning and end of (sub)turn-units.

As a recurring structural example, the phrase, e.g. /se paraka/lo/ (Fig. 1a) appears with a pitch gesture on the initial (unstressed) syllable, in addition to the stressed syllable, in the second phrase /ja na 'bume 'torm/, apart from an initial pitch-gesture, has a widened pitch range, as a reinforcement of this (new) part of discourse. The next figure (1b) also exhibits an initial pitch gesture which is completed within the phrase /li'pon/; should this pitch gesture carry a lexical distinction rather than a discourse cue, the result would be *'li'pmon/, i.e. a non-existent word in standard Greek. For this initial pitch gesture which, regardless of the rhythm-system status of the syllable, appears with a discourse function to attract the listener’s attention toward a particular unit of speech, we propose the term “initiative accent.”

In contrast to initiative accent, pitch-gestures may appear at the end of a (sub)turn-unit with distinct discourse functions. The phrase /ena ber/matio xarto/tilaka/ (Fig. 2a) carries a final accent which is realized as a pitch-fall on the last stressed syllable. This accent signifies the end of a sub-turn-unit and the completion but not necessarily the end of the ongoing turn-unit, and we may refer to it in want of a better term, as ‘completive accent’. On the other hand, at the end of a sub-turn-unit (Fig. 2b), the final (unstressed) syllable of the word /fiski/ carries a pitch gesture. This (upward) final accent is realized on the last syllable(s) of a (sub)turn-unit rather than the last stressed syllable. It has a turn-keeping function, but it may also be used as an ‘expectative’ discourse cue (expecting some response from the hearer) when addressing the listener(s). As a cover term we may use the term “continuative accent.”

A final accent may also appear at the end of a (sub)turn-unit associated with what has traditionally been called a question. Without going into an argument of what a ‘question’ is (see [3]), we present four wh-questions with two typical intonation patterns (Fig. 3). The first two (3a, 3b) both have falling final intonation but different communicative functions: (3a) is a pseudo-question, where the speaker is not sure, but the listener(s) must try to guess the question; (3b) is a ‘neutral’ question, i.e. the answer is of limited importance to the speaker and/or the development of discourse. On the other hand, the second two questions (3c, 3d) have a complex falling-rising intonational pattern, in which the final pitch gesture is co-ordinated with the final rather than the stressed syllable: these questions, the intonational pattern of which is very regular in Greek, may add a new dimension in the sense that the accent is of vital importance to the development of the discourse and, in this particular case, the outcome of the question.

The final pitch gestures, either for questions or continuations are quite similar in manifestation and partly share the same function, namely the emphasis put by the speaker in the development of the discourse. Of course, they are the speaker’s conditions because, in real life communication, he may get no answer or may be interrupted. Thus, preliminarily, we may use the term continuative accent in a ‘more to come’ broad sense even for emphatic questions, with the assumption that earlier prosodic cues and the context may distinguish them from turn-keeping pitch gestures.
at a low pitch level; an adult female finishes her phrase /pîn'lojî ti of'leans/ (Fig. 4b) at a low level and her interlocutor, an adult female also, responds /'ve vois/, with a pitch contour at the same level, in accordance with her communicative agreement. This by no means implies that the communicative distinction of agreement – disagreement is carried out solely by prosody; the lexical and grammatical components may be largely decisive. Nevertheless, our data have shown a pitch-concord in rather absolute terms than relative ones between different speakers. It seems that when a speaker chooses to indicate his agreement by prosodic means, he makes an extra effort to approach the actual pitch contour as close as possible.

3.2. Global Intonation

An interesting question is how speakers organize their overall intonation in terms of pitch range for discourse purposes and what the interference of external conditions like sex, age, etc. are.

Although in a more comfortable conversation [5] we have found pitch range as a turning point and negotiating discourse correlate, in the present material this phenomenon is drastically reduced. In other words, in a vivid interaction speakers seem to take advantage of e.g. the presence of the complete accent or even the absence of a turn-taking accent to intervene rather than using the pitch range-turn-taking cue. The same strategy is generally applied for topic management as well, in combination with the communicative context which appears as an everywhere factor for topic regulation. This reduces the potential of pitch range as a discourse mechanism for the government of turn/topic regulation, which is only occasionally realized in this kind of quick dialogues but is used at the end of the whole conversation.

On the other hand, a pitch range accent (prosis) reflects the involvement of the speaker(s) towards what it is said, i.e. it may span a succession of sub-turn-units and even have an interpersonal effect. Pitch range may also indicate focus, although a (major) pitch fall in combination with a post-focal accentless rhythmic organization is the rule in Greek intonation. However this regular manifestation does not leave the notion of focus unproblematic. As a matter of fact, in our material, we have witnessed only a few occurrences of focus, even in an auditory analysis. This indicates that focus is optional even for larger discourse domains such as turn-unit and topic, and not a recurrent prosodic category at a certain linguistic or discourse level.

Obviously, what speech analysts have described as 'focus' needs re-evaluation in a discourse perspective.

As regards the overall inter-speaker pitch range adjustment, our data has hardly shown any interference of sex or age. A preliminary evaluation shows that speakers do not mutually modify their pitch range but rather retain their idiosyncratic intonation except in cases of pitch-concord (cf. Fig. 4) where speakers choose prosodic means to show their communicative agreement. In more private and/or intimate communicative environments, another picture may arise, but this is a subject outside our current research.

4. CONCLUSION

In prosodic research we have experienced in the laboratory of the question – answer paradigm, where the answer is a declarative utterance making up the test material, the distribution of pitch gestures is clear-cut: the stressed syllables may appear with an independent (upward) pitch gesture whereas the unstressed ones either they have no inflection or they carry on a pitch gesture already started on the stressed syllable [1]. This neat picture is heavily disturbed in spontaneous speech where (upward) pitch gestures may appear on unstressed syllables and downward ones on stressed syllables. However, a closer examination reveals that this apparently contradictory prosodic manifestation has a meaningful structure.

Unstressed syllables with an upward pitch gesture may appear at the beginning (initiative accent) or end (continuative accent) of prosodically coherent larger units regardless of the stressed – unstressed distribution. Moreover, stressed syllables with neutralized pitch have a high rate in these Greek dialogues, solid evidence that pitch is not used to realize stress distinctions in Greek. Thus, rhythm and intonation appear quite independent, with intonation as a par excellence TOP-DOWN prosodic parameter, specifically meaningful in interaction and discourse communication.

BIBLIOGRAPHY


THE MOST IMPORTANT DIFFICULTIES WHEN TEACHING SPANISH PHONETICS TO CZECH

Jana Kullová
Fac. of Philosophy, Charles University, Prague

ABSTRACT

The most important difficulties when teaching Spanish phonetics to Czech native speakers are closely related to the rhythmic segmentation of the Spanish utterance. The differences between the two languages are relevant both on the segmental and the suprasegmental levels and become important not only from the point of view of the production of the speech signal, but also from the point of view of its perception.

1. INTRODUCTION

The aim of our paper is to point out some problems which students of Spanish, whose mother tongue is Czech, grapple with. We will pay our attention to the problems linked up with different continuous speech segmentation into rhythmic units in Spanish and in Czech.

The problem can be seen on two levels: a) speech production; b) speech perception.

2. SPEECH PRODUCTION

When analyzing the segmentation of the Spanish continuous speech into rhythmic units, among the sound means, the suprasegmental phenomena are almost exclusively taken into account. The rhythmic units are, as a rule, defined exclusively on the basis of only suprasegmental means conceived abstractly regardless of their concrete realizations in the flow of the speech, and on defining the rhythmic unit, exclusively one suprasegmental phenomenon is often taken into consideration.

If the phenomena concerning the definition of the rhythmic unit delimitation are taken into account, only the pauses, which stand, in fact, outside the rhythmic unit itself, are considered. Therefore the so defined rhythmic units become more likely a theoretical construct serving for the language description and only seldom represents the unit being perceived like that by listener.

On the other hand, when defining the rhythmic unit as a rhythmic semantic group, it must be conceived as a sound unit corresponding to a grammatical and a semantic unit, whose sound boundaries are marked by an interruption of the flow of speech potentially realized by a pause, by differences in the distribution of the positional variants of voiced consonantal phonemes, by glottal stop, by other sound phenomena or their combinations which, at the same time, carry suprasegmental means (stress, intonation, quantity) functioning, in accordance with the role the rhythmic unit plays within the levels of the structure of the utterance, as modulations of connected speech.

On the basis of the analysis of single features of the so defined rhythmic units, we consider the glottal stop in Spanish to some partial aspects which may cause difficulties in teaching Spanish as second language on the basis of Czech as the mother tongue to be described.

2.1. Rhythmic-semantic group delimitations

Important features differing Spanish from Czech are the differences in the distribution and the units of voiced consonantal phonemes /b/, /d/, /g/. These phonemes present in Spanish the occlusive + b, d, g and the fricative variants + , .

The distribution of these differs from one another in accordance with their position in the rhythmic sematic group: at the beginning and inside the unit after nasals the occlusive variants are used; the fricative variants appear in other positions.

There is thus a difference in the pronunciation of Casa de la and de Casa [déska], Barcelona [barcelona] and de Barcelona [debarcelona], etc. Czech native speakers do not recognize this phenomenon and often pronounce the occlusive variants [c], [d], in both positions.

Other phenomenon which is connected with the problem of delimitations of the rhythmic unit is the glottal stop. Considering that in Czech the glottal stop occurs automatically at the beginning of utterance if the first phone is a vowel, and the literary pronunciation requires the glottal stop after non-syllabic prepositions and in other cases the pronunciation of the glottal stop is motivated phonostylistically, Czech native speakers try to transfer this pronunciation into their speech in all these positions into Spanish. Instead of en air en ril they pronounce en a ril.

2.2. Syllable structure of connected speech

Other problems related to the syllable structure of connected speech are closely linked up with the above mentioned problem of distribution of the occlusive variants of voiced consonantal phonemes.

When analyzing the syllable structure within the rhythmic unit in Spanish, we find that the sound coherence of the rhythmic-semantic group determines its division into syllables, which is a phonemic process and there is no difference between the syllables or the rhythmic units.

At the beginning and inside the unit after nasals, the occlusive variants are used; the fricative variants appear in other positions.

There is thus a difference in the pronunciation of Casa de la and de Casa [déska], Barcelona [barcelona] and de Barcelona [debarcelona], etc. Czech native speakers do not recognize this phenomenon and often pronounce the occlusive variants [c], [d], in both positions.

Other phenomenon which is connected with the problem of delimitations of the rhythmic unit is the glottal stop. Considering that in Czech the glottal stop occurs automatically at the beginning of utterance if the first phone is a vowel, and the literary pronunciation requires the glottal stop after non-syllabic prepositions and in other cases the pronunciation of the glottal stop is motivated phonostylistically, Czech native speakers try to transfer this pronunciation into their speech in all these positions into Spanish. Instead of en air en ril they pronounce en a ril.

Besides the importance of the syllable as a component of the rhythmic-semantic group, we consider necessary to mention above all one of the features of the Spanish syllable: the tendency to

290

291
its openness. One of the manifestations of this phenomenon is the superority of the syllable structure to the lexical one within the scope of the rhythmical-semantic group, as mentioned above, and also several assimilation phenomena become very important. As for the type and the direction of assimilation, the articulation assimilation occurs more frequently, especially as for the place of articulation. The unstability of the place of articulation of nasals and laterals may be considered as a manifestation of this fonosyntactic phenomenon: con todo [kontudo] assimilation of the place of articulation, etc.

In the Czech language, the situation is rather different. A fundamental type of assimilation is the assimilation of voice. Owing to these differences, the Czech native speakers a) do not respect the assimilation of the place of articulation in Spanish; b) pronounce the consonants with the assimilation of voice.

Other problem of the syllable structure of connected speech is closely related to the above mentioned glottal stop, because of its absence in Spanish, due to the phonosyntactic phenomenon called synalepha. It means that the Czech native speakers do not avoid the sound pronunciation of expressions like a Ana [a:na] with the glottal stop [a:'ana].

2.3. Stress
Further problems are linked up with the word stress within the scope of the rhythmical-semantic group. Unlike the Czech, the quantity has a determinative function, and has a definite quantity, Spanish is a language where the stress falls on different syllables, considered as that of a given word category, and therefore has a distinctive function. On the other hand, not all "distinctive" stresses are realized with the same intensity. In the flow of the speech pattern, even stressed syllables which do not carry the distinctive stress (so called unstressed words - conjunctions, prepositions, unstressed forms of personal pronouns, etc.) In these cases, the stress is consideredconstantly and it is realized within the scope of the rhythmical-semantic group.

2.5. Quantity
The problem of quantity is also closely linked up with the problem of stress. If we start from the statement of incompatibility of free stress and phonological quantity, we find that the difference between Spanish and Czech consists in the fact that the quantity is phonological in the Czech language, while the quantity (duration) is sometimes closely linked up with the stress position. But considering the quantity as a sound parameter of connected speech, we find that the relation between both studied languages seems to be more complex. As a consequence of synalepha, or it can be considered also regarding the position of the respective syllable with reference to the stress.

The fact that the quantity has phonological validity in Spanish often causes that Czech speakers do not respect differences in Spanish vowels in different positions.

3. REMARKS ON THE SPEECH PERCEPTION
When analyzing the problem of the Spanish fluent speech perception by Czech native speakers, we must deal with difficulties caused mainly by two features of the above mentioned rhythmical-semantic group, both related with the syllable structure within it: by synalepha and by the assimilation phenomena. Both phenomena complicate the determination of the lexical units as components of the rhythmical-semantic group, and therefore the comprehension of its sense.

4. CONCLUSIONS
When summarizing the notes concerning the aspects defining the rhythmical-semantic unit in Spanish from the Czech native speakers point of view, it can be seen that the selection of sound qualities of the rhythmical-semantic unit is the starting point for doing analysis of an inadequate pronunciation of Spanish as foreign language, and it enables to find a common denominator for interpretation of a number of sound phenomena which would be otherwise correlated with difficulty. The emphasis on understanding of sound relation within the rhythmical-semantic group is important not only for explanation and training of the correct pronunciation of suprasegmental means, but it also enables a more profound view even on relations between segmental means, e. g. where a mere comparison of articulatory and acoustic features and repertory of consonants in Spanish and Czech, differences in assimilation, etc. is not sufficient.

5. REFERENCES
/3/ KULLOVA, J. (1968), "Mod úvazon雅 de la cadena hablada en español", Praga: AUC.
ASPECTS OF THE RELATION BETWEEN INTONATION AND THE INTERPRETATION OF POEMS

Johann Z. Uys

University of Fort Hare

ABSTRACT

Two hypotheses concerning the relation between intonation and the interpretation of poems were tested: firstly, that appropriate renderings of poems could contribute to a closer indication of 'possible' (and probable) meanings, and secondly, that instances of diverse interpretations could occur when individuals (including the poets themselves), render poems in accordance with their own personal opinions.

1. INTRODUCTION

This paper is intended as a contribution towards the illumination of the relationship between the intonation of poems and their interpretation. Poets and literary critics alike generally claim that the sound structure of poetry is important. Too often, however, interpreters of poems only pay lip-service to this fact. Although some attention is paid to sound phenomena such as alliteration, assonance and rhyme, these are static aspects that are determined by the lexical structure. The author is not aware that the dynamic aspects of poetry have been investigated systematically with a view to establishing their contribution towards the overall meaning and impact of poems.

The hypothesis presented here is that an exhaustive interpretation of a poem requires all possible renderings of the poem to be taken into consideration. The reduced hypothesis is that any one interpretation rests to an appreciable extent, on the intonational dynamics of a particular rendering. In general the particular rendering is in the mind of the interpreter and he/she does not make explicit its particular structure. Thus, the contribution of the particular dynamic structure of the intonation remains hidden, and the difference in interpretation between two persons 'imagined' rendering remains unexplainable.

2. METHOD

Several mother-tongue speakers of Afrikaans were asked to recite a number of Afrikaans poems. These were recorded on tape in a professional recording studio. A group of 20 mother-tongue listeners was then asked to determine the acceptability of these renderings of the poems on a ten point scale. This procedure led to two poems being selected by all subjects as having been rendered adequately in all respects. These two poems, "Skuiling" and "Sproetseena", are both by D.J. Opperman.

The two poems were then analysed acoustically, focussing on the extraction of the Fo contours. Of the one poem, a recording by the poet himself was available on cassette tape, but because the quality of the sound-track was poor, the mother-tongue speaker who had recited the other two poems, was requested to imitate the poet's own rendering as closely as possible. This was analysed in the same way as the other two, utilizing the equipment and Fo extraction programme of the Institute of Perception Research of the University of Technology, Eindhoven (Netherlands). (Cf. Hermès 1988).

3. RESULTS

A print-out of the Fo contour of the four lines of poem no. 1 ("Skuiling"), clearly revealing which words are receiving prominence through increased pitch, is provided in Fig. 1.

Fig. 1 The four lines of the poem "Skuiling" compressed into two run-on lines because of enjambment.

Fig. 2 and 3 represent the versions by a mother-tongue speaker and by the poet himself of the particular line indicated, viz. "..weet ek hoe dat 'n vrou kan troos"
Fig.

Fig. 2 The realization of the line "...weet ek hoe dat 'n vrou kan troos" by a mother-tongue speaker.

Fig. 3 The realization of the line "...weet ek hoe dat 'n vrou kan troos" by the poet himself.

("... do I know how a woman can comfort"). This line has been selected, because it exemplifies a marked difference in accentuation and intonation.

4. DISCUSSION

The poem "Skulling" (Eng. "Shelter") has been selected because the interpretation of this quatrain has been outlined clearly in literary criticisms (cf. Scholtz 1978: 102). According to these views, the unborn child is addressed and advised that, although it still finds safe shelter in its mother’s womb provisionally, it will realize soon that we human-beings of skin and bone, are very fragile. Now, the acoustic realization of this poem does not alter the overall meaning of the poem as such, but it does seem to focus particular attention to certain "propositions". These propositions all happen to be words loaded with modality, viz. the adverbs "voorlopig" ("provisionally"), "veilig" ("safety") and "ook" ("also") and the adjective "nietig" ("fragile").

The relatively “simple” interpretation of the poem should, therefore, be relativized. The strong reliance on adverbs and adjectives lend a particular modal ‘colour’ to the otherwise straight-forward interpretation. (cf. Oakeshott-Taylor, 1984.)

Turning to the second poem, the selected portion illustrates the dependence of one interpretation rather than another on a particular realization and underlines how extremely useful it is to have a rendering by the poet himself. Figs. 2 and 3 show the interesting contrasts that can be created by comparing the mother-tongue speaker with the poet himself.

Within the same rhythmic structure, different locations of tonal accent shift the focus of the line from “weet” (Eng. “know”) (mother-tongue speaker) to “vrou” (Eng. “woman”) (poet) (cf. Crutenden, 1986: 89).

From the orthographic form of the poem, both interpretations are latent, but the realization dynamics make only one or the other possible.

5. CONCLUSION

Both hypotheses tested were confirmed, viz.

1) that the specific realizations of the poems at hand, focussed special attention to certain key propositions, thereby providing more concrete substance to the illocutionary force of the message, and narrowing the field of alternative interpretations;

2) that two different renderings of a poem reveal ever so slight, but highly interesting differences in emphases.

The overall conclusion that seems warranted by the result, is that the intonation pattern of a poem does have important implications for the interpretation of such literary works.

6. ACKNOWLEDGEMENTS

The author wishes to thank the Institute of Perception Research of the University of Technology, Eindhoven, more particularly Dr. Leo Vogten, for enabling him to do the necessary acoustic measurements. A warm word of thanks is also due to Dr. Bill Barry of University College London, for invaluable assistance in the formulation of the final text.

7. REFERENCES

EFFECTS OF VOICE CHARACTERISTICS ON ATTITUDE CHANGE
C. Gélinas-Chebat and J.-C. Chebat
University of Quebec at Montreal

ABSTRACT
A 2 x 2 x 2 factorial designed experiment (2 levels of intensity x 2 levels of intonation and 2 levels of involvement) showed that these two voice characteristics play the role of credibility in the Elaboration Likelihood Model. Main effects of both prosodic characteristics and combined effects of these two characteristics on receivers' attitude toward the message prove to be significant only under low involvement: low intensity and low intonation enhance attitude change, as high credibility does.

1. INTRODUCTION
Very few psychosocial studies investigated the prosodic antecedents of credibility (e.g. Page and Balloon, 1978); some other, more frequent, phonetic studies have investigated the mental image derived from the speaker's voice (e.g. Brooke and Hung Ng, 1986). No research seems to have bridged the gap between the two research areas: this study shows how two prosodic characteristics can be integrated in a widely accepted psychosocial model, the elaboration likelihood model (ELM).

2. ELABORATION LIKELIHOOD MODEL
Petty and Cacioppo developed ELM to explain attitude changes: briefly, they mapped two basic routes of persuasion: A central route which occurs when the person is motivated and able to think about the issue and a peripheral route which occurs when either motivation or ability is low (1981:365). The central route is followed when message arguments enhance the cognitive justification of (...) issue relevant information" (Petty, Cacioppo and Schumann, 1983:135). The peripheral route is followed because the issue is associated with positive or negative cues or because (...) simple cues in the persuasion context" (Petty, Cacioppo and Schumann, 1983:135). Among these cues, speaker's credibility constitutes a major one. To the extent that one possesses only a limited amount of information processing time and capacity, the fact of scrutinizing the plethora of counterattitudinal messages received daily would disengage from the exigencies of daily life. ELM proposes a principle of information-processing parsimony according to which consumers seek to process as little data as necessary.

However, no study so far seems to have investigated the voice cues in terms of antecedents of credibility within a structured psychosocial model of attitude change; this is the purpose of our study.

3. PROSODIC CHARACTERISTICS IN SOCIAL PSYCHOLOGY
Prosodic characteristics were studied as indicators of speaker's emotion (e.g. Fonagy, 1983; Leont, 1971) or speaker's personality (e.g. Berger and Kellerman, 1989) or speaker's social status (Pittman and Gallois, 1986) or speaker's persuasive capacities (Brooke and Hung Ng, 1986) or arguments plausibility (Ekman, 1988). Hall (1989) showed that speakers' persuasibility depended on their manipulated voice characteristics: some specific voices, perceived as "warm", "expressive" or "calm", proved to enhance speaker's persuasibility. The reviewed literature is showing two kinds of studies: in one hand some studies show the voice antecedents of credibility without showing their effects on attitude; on the other hand other studies show the effects of perceived voice on persuasion without pointing at the prosodic causes of these effects.

4. METHODOLOGY
A 2 x 2 x 2 factorial design was used: 2 levels of issue involvement x 2 levels of intonation x 2 levels of intensity. Two mock advertising messages, the linguistic characteristics of which were as close as possible to each other, were designed for a public of business students: the topic of the low involve-
ment advertising message was the ATM card; the topic of the high involvement advertising message was the students' loan. A professional comedian was instructed by the first author to manipulate his voice to produce high versus low intensity and low versus high intonation. A group of 30 linguistics students was used as judges to assess the prosodic variations. A questionnaire on attitudes toward one of the two financial services advertised was administered to eight approximately equal groups (total N = 279) of business students of our university. 221 questionnaires were completed and usable. Manipulation checks showed that, for the tree dimensions of the factorial design, the low level was significantly different from the high level counterpart.

5. RESULTS
An analysis of variance, (the dependent variable of which is the attitude toward the advertised service), shows that: Neither information nor intensity has main effects on the dependant variab-

As predicted by ELM, issue involvement has significant main effects (F = 14.37, p = .000).

As predicted by ELM, however, both intonation and intensity significantly interact with involvement (F = 3.21; p = .075 for intonation and F = 2.98; p = .086 for intensity), see Fig. 1 and Fig. 2.

As predicted by ELM, a three-way interaction between intonation, intensity and involvement significantly interact (F = 5.000; p = 0.026). See Fig. 3a and 3b.

Unexpectedly, a two-way interaction between intensity and intonation is found significant (F = 3.21; p = .075), see Fig. 4. However, when the receivers' involvement score (Zelkowsky, 1985) is held as a covariate these interactive effects are no more significant, (F = 1.494; p = .223).

6. DISCUSSION
ELM is basically confirmed. "Peripheral" prosodic cues have significant effect only under low involvement. More precisely, low intensity and low intonation as well as the combination of low intensity and low intonation prove to produce higher attitudinal scores than the high counter parts. Hall (1979) found that in some specific cases "more stiff and less warm" voices produced better persuasive effects. In the absence of other similar studies, we reason that high profile speakers could enhance receivers' defensive mechanisms which are attenuated under low involvement.

Our study is confirmatory of some European phonetic studies by Goldbeck et al. (1988) who showed that these are "interactions between (intonation) contour and text in communicating aspects of speakers' affect" (p. 129). Our study shows that the low involvement text enhances the effects of prosodic characteristics which play the role of credibility in ELM.
References


PAROLE CHANTEE ET PAROLE DECLAMEE: AUTOUR DE SALOMÈ
ASPECTS ARTICULATOIRES, RHYMHIQUES ET INTONATIFS

Sibylle Vater
Centre de phonétique expérimentale
Université de Genève, Suisse

ABSTRACT
There are a German and a French version of the opera Salomè, and there is a Marjolitte's opera. This complex makes excellent material not only for comparing the vocal lines of two contrasting languages and relating them to Marjolitte's vocal conception, but also for analyzing fundamental differences between spoken word and word performed by singing.

1. INTRODUCTION
Le chant, qu'il soit soutenu par l'orchestre ou non, révèle et libère certains traits de la parole qui, dans la déclamation, restent jugulés par les contraintes du code phonétique et syntaxique. Chaque compositeur s'y prendra à sa manière et en fonction de l'âge du livret. Cependant, afin de se mettre au diapason du mètre, le chant doit prêter à ce dernier ses propres structures, notamment celles d'ordre rythmique et tonal, celles donc que le linguiste qualifie de supra-segmentales.

2. SITUER LE SUJET
Notre étude cristallise autour du complexe de Salomè qui, au début du siècle, s'est formé à partir du drame d'O.Wilde(OW). D'abord, il s'agit des deux versions de l'opéra de R.Strauss(RS) dont la première (1905) repose sur la traduction allemande de la pièce française par H.Lachmann(HL) et la seconde (1906) directement sur cet original. En effet, dès l'achèvement de son opéra allemand, le compositeur désirait rendre justice au texte français, ce qui l'engage à adapter de nombreuses lettres, de mot en trouvant un tonage restant inchangé. Pour réussir son entreprise, RS demande à R.Rolland(RR) des conseils en prosodie française [5]. De son côté, Jean de Marilave traduit en français la traduction allemande de HL tout en se conformant à la ligne vocale allemande (1909). De plus, toujours à la même époque et d'après la pièce d'OW, A.Marijotte (AM) compose son opéra Salomè.

3. OBJECTIFS
Dans le cadre restreint du présent exposé, nous nous limitons à l'analyse de quelques exemples correspondant à de la -la version allemande de RS(Rs)-, la version française de RS(RSF), Salomè d'AM.

D'une part nous comparons la parole chantée à la parole déclamée, d'autre part nous examinons entre elles les lignes vocales des deux œuvres. Plusieurs échantillons seront étudiés en fonction de l'articulation dramatique de l'opéra.

4. MÉTHODE EXPÉRIMENTALE
Les séquences déclamées ainsi que les paroles rythmées et chantées seront analysées à l'aide d'oscillogrammes, d'intensigrammes et de courbes intonatives. A cette fin il faut accéder au phrasé et au chant pur, non soutenu par l'orchestre et développé de tout bruit technique. Nous visionnons des interprétations qui respectent l'écriture musicale dans son ensemble et nous remarquons N.d.davoir bien voulu les assumer. Pourtant, qu'on ne nous tienne pas rigide d'exécution approchée de certaines temps et intervalles. L'essentiel n'est pas là. Les points que RS et RR soulèvent dans leur correspondance restent en marge de notre optique.

5. JUSTIFICATION DES EXEMPLES
La pièce d'OW abandonne en symboles et phrases récurrents qui concourent à charpenter le drame et à créer un vertige, tout en se transmettant au bénitier d'un personnage à un autre. RS et AM - ce dernier dans une monodie - reprennent ces leitmotives à leur tour, par exemple:

- RS(F/AM Page): Vous la regardiez toujours. Vous la regardez trop; - RS(Salomè): Narraboth, je vous regarde... N. regarder-moi... (chez AM presque pareil).

Précisément, RS crée un "Musikdrama" où de nombreuses séquences sont intimement liées au resserrement de l'action. Trois escaladons se profilent en particulier.

Nous en retiendrons ici celle qui entraîne l'exigence extrême de Salomè (scène 4).

6. PRÉSENTATION D'ANALYSES
6.1. L'exclamation initiale Narraboth:

- RS: "Wie schön ist die Princesse Salomè heute noch!

- RSF: "Oh, oh, ohne governmental vous vais rien faire, Amie..."

- AM: "Oui, je veux que l'on m'apporte maintenant, dans un bassin d'argent..."

La tête d'Okanan.

- AM: "Je veux que l'on m'apporte maintenant, dans un bassin d'argent... la tête d'Okanan..."

Grâce à un développement oratoire intermittent de Hérode, OW et RS entretiennent habilement le suspens par lequel s'interrompt la séquence initiale. Puis l'exigence de S tombant comme un coup de tonnerre.

Notamment dans la version française, RS réussit une symphonie parfaite entre la parole et la composition musicale (chant et orchestre). Comparée avec le texte allemand qui conclut le paroxysme de l'épisode d'OW, la rhésie introduit une percée, dépouillée de tout verbe, paradoxe remanié. Ainsi la préparation du suspense s'intensifie. Quant à la modulations qui, dans les deux versions, affecte...
le nom du prophète, surtout le traitement rythmique, tonal et articulatoire du premier a démontré un éclatement de potentiel phonémique suprême (v. 111.4).

Au contraire, par la suppression de la parenthèse d'Hérode et par une ligne vocale plate, AM compromet gravement l'effet dramatique (v. 111.5).

7. CONCLUSION
Dans la parole chantée, le rythme et la graduation tonale permettent un grossissement maximal du centre syllabique. Soumise à des contraintes fonctionnelles incontournables, le langage parlé ne peut pas accéder à des dilatations pareilles; cependant, il possède d'autres ressources, celles de la poétique, par exemple.
PHONOStYLISTICS IN FOREIGN LANGUAGE LEARNING

Kathryn G. Keller
Summer Institute of Linguistics, Mexico

ABSTRACT

In learning to speak a foreign language with as little mother-tongue interference as possible, the student needs to be able to recognize and control phonetic features which occur over strings of speech, both those which characterize the language as a whole and give it its distinctive character, and those which occur contrastively within the language to express affective meanings. Illustrations are presented from a number of American Indian languages.

1. INTRODUCTION

The language learner should approach the pronunciation of the target language from two points. It is important to be able to pronounce the individual sounds and be able to put them together into words and sentences. It is important also to tackle language from the other end, starting with longer utterances and paying attention to overall phonetic features such as rhythm, stress of utterance, pitch patterns, loudness, tongue position, lip shape. There are two areas in which phonetic features occurring over strings of speech are important: those which characterize the language as a whole and make it sound different from another with which I'll call the overall features of a language and those which are stylistically contrastive within the language. The term phonostylistics is used here to cover both areas.

The illustrations cited have been gathered over the years in personal conversations with SIL colleagues and in response to a questionnaire circulated among them. (Due to space limitations, I cannot list them individually.)

2. OVERALL FEATURES

Beatrice Honkiman [1] has emphasized inherent differences in languages and the need to adapt the speech apparatus to the movements characteristic of the target language, that is, to shift gears, illustrating primarily from Indo-European languages. The principle of shifting gears can be profitably applied in learning to speak American Indian languages as well.

Atlatlahuca Mixtec (Mexico) is characterized by tongue frontedness. It is easier for a learner of that language, who comes from English as his mother tongue, to shift gears—move the whole tongue farther front in the mouth—than to remember each time he comes to individual sounds such as u, i, e, a, u that they must be farther front in the mouth than the similar sounds in English. Spanish is also characterized by tongue frontedness. A clue to general tongue position in a language is the hesitation forms. Spanish speakers hesitate on e e or este este in contrast to English speakers' est ya. As to rhythm, both Atlatlahuca Mixtec and Spanish exhibit syllable timing rather than stress timing as do English and Southern Tepehua.

Seminole (United States) is another language characterized by tongue frontedness, plus the feature of spread lips. There is very little jaw action except when the people are excited or when they are trying to speak precisely and exactly to a stranger they think wouldn't understand otherwise.

Various Indian languages are characterized by soft spoken speech. Among them are Comaltepec Chiantec, Yatzachi Zapotec, Atlatlahuca Mixtec, Eastern Popoloca, Seminole and Mazatec. The Seminoles speak so quietly that sometimes they are barely audible. This is in contrast to Tabasco Chontal (Mayan) and Veracruz Tepehuá where people generally do not speak softly.

The Mazatecs speak quietly. Women never raise their voices. In Huatlta there is a large market, full of hundreds of people, but you cannot hear those voices across a half a block away. If you do hear loudness, it is a drunk, a Spanish speaking person, or someone in a fight.

Some Mexican Indian languages have pitch downdrift, including the tone languages Tepetotutla Chianantec, Chihiualtan Mazatec, Coutzspan Mixtec, Quiquotec Chinantec (over a breath segment), and Yatzachi Zapotec (within phrases and clauses). Mura-Piraha (Brazil), on the other hand, may exhibit updrift of voice over a sentence. Its many glottal stops make it sound choppy.

The ballistic and controlled syllables of Amuzgo (Mexico) give it a distinctive rhythm. Kenneth Fike has described differences between four Peruvian languages in terms of ballistic and controlled abdominal pulse types [3]: Arabela, Culina, Aguaruna and Campa.

To learn to speak well, one needs to be aware of what overall features characterize a particular language. Listening over and over to connected speech on tape early in the language learning process increases awareness of these features.

Along with repeated listening to a text, the student should begin tracking, that is, speaking along with the tape as simultaneously as possible, not concerned about missing some segments, but aiming to reproduce the overall rhythm and pitch patterns, up to speed. A person can track silently whenever he hears the language spoken and he himself is not in focus, that is, not being expected to listen and respond. This will help fix the sentence melodies in his mind.

3. STYLISTICALLY CONTRASTIVE FEATURES

In addition to features which color a language as a whole, phonetic features occur within languages over strings of speech and are stylistically contrastive. These phonetic variations are socially significant, carrying meanings related to moods and emotions. Features such as height of pitch, width of pitch intervals, intensity, rate of speech, creaky voice, breathy voice and lip shape are sometimes referred to under voice quality [2] or prosodies, or as subsegmental features [4].

The language learner needs to be aware of the phonostylistic features in the target language in order to understand nuances of the spoken speech, and to avoid being misunderstood, insulted, or impolite when speaking.

John Crawford reports that when he lived among the Mize people, he could almost tell when a visitor was leading up to asking to borrow money, as the visitor always used creaky voice. A mad, excited Mize speaker used a monotone with a dive down at the end. For emphasis or excitement, the
speech was breathy.
In Huatla Mazatec anger is shown by lengthening the vowels, not by raising the pitch as may occur in American English. A Mazatec child, wanting to look at a book that another child has had for too long, may say (translation): 'It's mine, turn it over.'

Urgency, on the other hand, is expressed by breathiness, as when impatiently calling for someone: 'Victor, Victorjishah!' Sympathy is shown by lip rounding accompanied by poked out lips.

3.1. Differences in Feature Use From Language To Language
The language learner needs to be aware that the same phonetic feature may signal different things in different languages. For instance, lip rounding in Quiché (Guatemala) indicates a compliment. In some Mazatec and Mixtec languages (Mexico) the lip rounding, accompanied by poked out lips, is used in showing sympathy. In Zuni (southwestern United States), lip rounding accompanied by poked out lips and low pitch, is used for scolding, as when a father says to his son, 'You’re just a one feather Indian.'

3.2. Some Common Meanings Expressed Phonostylistically
3.2.1. Scolding Children
For scolding children, a frequently used feature is higher pitch. The high pitch is accompanied by loudness in Highland Chontal, Jalapa de Diaz Mazatec, Alacatlaltiza Mixtec, and Ocotlan Zapotec. The high pitch is sustained in Highland Totonac: without lowering. In Caucatec and Cora (Mexico) it is accompanied by fast speech. In Cora the pitch is so high it is almost falsetto, and the rapid speech has few final pauses.

Languages for which lowered pitch is reported are Western Ixtlan Zapotec, Northern Mixteco Mixtec, and Atalatla Mixtec. In each of these the speech is rapid, and with narrowed pitch range. In Western Ixtlan Zapotec the lips are somewhat pursed, and there is very little lip movement.

Lips are rounded and protruding in Yatzachi Zapotec. In Chatino the speech is very fast, and the tone contrasts are accentuated. In both Xicotepec Totonac and Comaltepec Chinante the speech is stacatto. In Chiuhuahua Mazatec there is exaggerated aspiration. Loudness, protruded syllables and some breathiness are reported for Ozumac Chinante. In Nahuatl of Teteelingo and of Orizaba there is an abrupt cutoff of phrases and sentences preceded by abrupt downturn of intonation. Mixtecan Nahuatl and Southern Tepehuán speakers talk quietly to their children. Tepetotula Chinante speakers use a "duckbill pout" (not rounded), with greater pitch spread, beginning high and ending low.

3.2.2. Talking to Babies
In talking to babies high pitch has been observed in more languages than low pitch. However, low pitch has been reported for Lacandon and Gueluva Zapotec.

Quite a few languages exhibit general fronting, or specific consonant changes such as palatalization. In Trique not only is there replacement of alveopalatalts by fronted alveopalatals or dentals, but sometimes replacement of dentals by alveopalatalts or fronted alveopalatalts. Atalatluca Mixtec /f/ is substituted for /j/, for /a/ and initial /s/ of consonant clusters is dropped. In Coatzogotlan Mixtec /s/ becomes /t/ in Veracruz Tepehuan the consonant changes are: /f/ > s /t/ > ts /t/ > q /q/ > k In Seri s > f.

3.2.3. Showing Sympathy
We have mentioned that lip rounding is used in Mazatec and Mixtec to show sympathy. In San Felipe Otomi and in Veracruz Tepehuan it is used both to express and elicit sympathy.
Creaky voice is reported for Alacatlaltiza Mixtec and Trique. In Trique falling pitch is superimposed on the tone system, and increased creaky voice occurs as the pitch falls; also the pitch at the end of the sentence is lengthened. Choapan Zapotec and Highland Oaxaca Chontal are softly spoken. Lacandon exhibits higher pitch and fronted tongue.

3.2.4. Showing Respect
High pitch, even sometimes falsetto, is used for showing respect in some languages. High pitch in Pame shows special respect to a comrade or comrade. Tenejapa Tzeltal women switch into a falsetto, along with averting their eyes when they want to show extreme respect, as to a person higher in rank, a town official or a witch doctor. The falsetto shows submissive attitude and sometimes fear. San Felipe Otomi speakers use falsetto to show politeness and respect. When compadres meet, for instance, they start out in falsetto, then drop back to ordinary speech as the conversation continues.

Falsetto is also used as a greeting for distance, or from outside the house when one comes to the house of a friend.

Another feature used is diminished volume. This softness is accompanied by more glottal stops utterance final in Jalapa de Diaz Mazatec, a language with all open syllables. In Alacatlaltiza Mixtec the soft spokenness is accompanied by lengthened vowels and rising-falling intonation on the last syllable of the words for respectful address occurring at the end of the sentence.

3.2.5. Anger
Anger is variously shown in different languages by high pitch, low pitch, rapid speech, slower speech, or sudden complete silence. There is also variation from wide pitch range to narrow pitch range. In Ozumacin Chinante the lower pitch is accompanied by lower volume. Tapanec exhibits short stacatto or nearly monotone utterances.

3.2.6. Asking a Favor
In Xicotepec Totonac the voice goes up and up if the speaker is about to ask a favor. Chiuhuahua Mazatec speakers, however, use lengthened vowels and exaggerated nasalization, which they also use when eliciting sympathy. Choapan Zapotec speakers are barely audible, with barely any mouth movement.

Falsetto is used in San Felipe Otomi when pleading for mercy. For example, a young boy being scolded and threatened with a whipping might switch into falsetto.

3.2.7. Emphasis
Heaver word stresses and wider pitch range were the most common features reported. In Mazahua there is labialization of consonants of the first syllable of roots, and sometimes lengthening of vowels. Zacatepec Mixtec exhibits word reduplication, vowel lengthening, and raised intonation. Consonants are more fortis in Jalapa de Diaz Mazatec. In Southern Tepehuan high pitch and lengthened vowels are used.

REFERENCES
RIEZ-VOUS EN HE! HI! HI! OU EN AH! AH! AH! OH! OH!

Pierre R.A. LÉON
Laboratoire de phonétique expérimentale
Université de Toronto, Canada

RÉSUMÉ
Laughter is an emotion. Its primary manifestation is physiological disorder, poorly structured, from an acoustical point of view. Socialized, laughter is less intense, with a more regular rhythm, more vocalized and even with a specific intonation. There, it becomes a signal in a semantic system. Its de-coding, as the one of emotions, depends upon contextual and individual factors. Tests show the greatest agreement on lexicalized laughter than on its real perception.

1. DU RIRE SÉMIOTIQUE AU RIRE SÉMANTISÉ

Une première recherche (Pierre Léon, Ron Davis et David Heap [5]) a permis de montrer en fait, au plan sémantique, l'existence de 3 grandes classes de rire (positifs, négatifs, indéterminés) tout en dégageant les tendances de leur structuration acoustique. Dans la présente étude, on a tenté d'étudier plus en détail le décodage du rire et les mécanismes qui sous-tendent son codage acoustique.

2. DÉCODAGE DU RIRE
On a administré un test de choix forcé sur les 10 étiquettes suivantes: masculin, féminin, amusé, joyeux, surpris, admiratif, colèreux, ironique, réprobateur, douloureux, autre. Le corpus était constitué de quinze rires, présentés à un groupe de vingt universitaires adultes francophones (dix hommes, dix femmes). On a obtenu les principaux résultats suivants:

- La différenciation rire masculin/féminin a été reconnue dans tous les cas sauf 3 exceptions: un rire très consonantique [kks]! (1 erreur sur 20); un rire féminin très intense, nettement timbré en [a] a été interprété comme masculin (2/20).
- 42% des rires ont été interprétés comme joyeux ou amusé.
- Dans les autres cas, la grande dispersion des réponses montre que le sémantisme attribué au rire, comme aux émotions, dépend beaucoup du contexte référentiel et situationnel. (Les rires entendus étaient hors contexte.)
- Seuls quelques rires ont été identifiés avec un accord relativement important: 2 rires ont été identifiés amusés à 40 et 50%; 3 joyeux à 50, 80 et 60%; 1 surpris à 60%; 2 ironiques à 40%, 40% et 60%; 1 généré et sexy à 40%; un âtre et sadique à 40%. Ces chiffres confirment bien l'existence d'un codage même si son fonctionnement reste souvent approximatif.

3. CODAGE ACOUSTIQUE
D'une manière générale, il semble difficile de tracer des limites acoustiques entre les diverses catégories de rires. On pourrait plutôt imaginer que les variables en cause, au lieu de former des classes discrètes, s'échelonnent graduellement sur une échelle allant du rire brut (cf. l'exemple de la figure 3) au rire conventionnel, (cf. l'exemple de la figure 1) de la manière suivante:

    brut ..........conventionnel

rythmicité = — +
intensité = + —
mélodie = — +
vocalité = — +

On va ainsi du désordre à l'ordre. Les pulsions dont le rire est fait sont toujours présentes bien qu'elles tendent à l'irregularité dans le rire brut.

Si l'on essaie maintenant d'examiner la structuration acoustique des rires dont on a donné ci-dessus l'identification sémantique, on relève quelques traits intéressants pour les échantillons analysés au mélomètre de Martin.

- masculin/féminin: l'opposition se fait essentiellement par le trait de hauteur, comme dans la voix.
- amusé et joyeux: semblent deux variétés; la première étant moins intense. Le rire joyeux est rythmé, fait de petites notes hautes, et bien timbré en [a]. On voit ainsi sur la figure 1, que F2 oscille entre 100 et 168 Hz, μ = 112, σ = 29 Hz. L'intensité générale est assez forte (32dB) mais les pulsions, très vocailes, ne sont séparées que par de faibles changement d'intensité (μ = 32.8 dB; σ = 2.2 dB).
- surpris: se manifeste ici par un souffle suivi d'une partie sonorisée, légèrement nasale et de montée mélodique rapide (fig.2), caractère du patron prosodique de la surprise.
On aspiration mélodique des Figure (entre celles quelques configurations acoustiques) et des sœurs d'intensité importantes (μ = 27.1 dB, σ = 8.4 dB).

Figure 4: Rire "géné et sexy".

- gêné et sexy: Le patron rythmique et mélodique est très irrégulier (fig.4). On entend beaucoup de souffle, une aspiration sonore forte et très aiguë sur la dernière pulsion (253 Hz) avec des sœurs d'intensité importantes (μ = 27.1 dB, σ = 8.4 dB).

- bête et sadique: Le patron rythmique commence par des pulsions lourdes (entre 20 et 25 cs) et se termine par une série de plus petites (8 à 10 cs). Le timbre est en [e] et la mélodie plate (μ = 122 Hz, σ = 8 Hz). (Fig 5)

Figure 5: Rire "bête et sadique".

On a pu constater ici, dans les quelques patrons analysés, des configurations acoustiques analogues à celles relevées pour les émotions dans la parole (Föna, 1983 [2]; Léon, 1976) [4].

4. L'IMAGINAIRE DU RIRE

On a retenu 6 graphies, qui nous ont parues, intuitivement, correspondre à des étiquettes attribuées au rire. Ces graphies étaient: [hi hi hi! ha ha ha! oh oh oh! hein hein! he hé hé! how how how!]. On a demandé alors au même groupe de 20 adultes de référer ces graphies à l'une ou plusieurs des catégories de rires suivantes: enfant, fille, garçon, joyeux, admiratif, réprobateur, sarcastique, douloureux, autres.

On a obtenu les résultats suivants: Entre parenthèses, le premier chiffre indique le nombre de réponses masculines, le second celui des réponses féminines. Le chiffre suivant donne le pourcentage. Les réponses inférieures à 10% (rares) ne sont pas indiquées ici:

Hi hi hi! enfant (6+6) 60%; fille (8+10) 90% joyeuses (4+4) 40%; sarcastique (2+2) 20%; Ha ha ha! garçon (6+10) 80%; joyeuses (6+6) 60% Ho ho ho! garçon (4+4) 40%; admiratif (8-2) 50%; réprobateur (4-6) 50%

Hein hein hein! réprobateur (4+0) 20%; sarcastique (8+8) 80% He hé hé: enfant (2+6) 40%; fille (2+4) 30%; sarcastique (6+4) 50% Hou hou hou! garçon (6+6) 60%; réprobateur (0+4) 20%; douloureux (6+6) 60%

Les réponses de la colonne autres ont été assez rares. On a relevé pour [hi hi] nerveux (15%); ironique (10%).

- On voit très bien se dessiner dans l'imaginaire des sujets parlants le rire en [hi hi] comme celui d'un enfant ou d'une fille, connotant ainsi le trait acoustique + aigu du [i] avec une voix naturellement haute; ce que confirme la notation de nervosité, venant du trait acoustique + tendu.

Le rire en Ha ha ha! n'est jamais attribué à un enfant ou à une fille, ce qui est infirmé par l'écoute quotidienne, tout au moins chez les femmes adultes. L'imaginaire se réduit au rire du garçon (80%), joyeux (60%). Et ce sont les auditrices (10 sur 16) qui ont été les plus nombreuses à voir là un rire essentiellement masculin, qualifié par quelques sujets de relax.

Le rire en Oh oh oh! n'a pas non plus été attribué aux filles, peut-être à cause du trait acoustique + grave, connoté avec les voix masculines. Il est intéressant de constater que les votes se partagent également entre les séries d'admiration 50% et de réprobation 50%. Il s'agit vraisemblablement, d'un côté, de la projection d'une voix haute, avec courbe exclamative et timbre clair, opposée à celle d'un ton grave avec timbre plus sombre.

Le rire en Hein hein! est jugé réprobateur (20%) ou sarcastique (80%) et également supérieur. Tous ces sèmes se rejoignent et confirment l'observation freudienne d'Ivan Föna (2) attribuant à la nasalité ces différentes connotations.

Le rire en Hé hé hé! n'est jamais attribué à un garçon mais à un enfant (40%) ou à une fille (30%). Il est en effet le trait acoustique + aigu du [e] joué comme pour le [i]. On constate alors que ce type de rire féminin est connoté avec les sèmes de sarcasme (50%) voire de méchanceté (15%) ou d'ironie (10%).

Le rire en Hou hou hou! n'est jamais attribué à une fille mais à un garçon (60%) avec les sèmes de réprobation (20%) ou de douleur (60%), provenant sans doute du trait acoustique + grave.

Le rire conventionnel, vocaliquement timbré, paraît ainsi receler un symbolisme très nettement codé dans l'imaginaire paralinguistique des sujets francophones testés. Il serait intéressant d'effectuer le même type d'enquête sur d'autres groupes linguistiques. Notons que le variable sexe n'a paru avoir ici qu'une très faible incidence.

6. CONCLUSION

La structuration acoustique du rire, son encodage, du sémiotique au sémantique, le placent bien dans la classe des émotions. On n'a examiné ici qu'une petite partie de saction identificatrice, indice des variables de sexe et d'émotion.

De l'indice, le rire passe au signal en se sémantisant. Ses diverses formes constituent alors un code dont les signes motivés sont néanmoins devenus suffisamment conventionnels pour fonctionner dans le processus d'une communication très spécifiquement humaine.

7. RÉFÉRENCES

VARIABLES INTONATIVES CHEZ LA FEMME VENEZUELLE

Elsa Mora

Universidad de los Andes, Mérida, Venezuela.

ABSTRACT
This paper analyses intonation in the registers of Venezuelan women of different social classes. Graphical representation of the results is done using the intonation model proposed by Fant (1984). Results show social class phonostylistic differences.

1. INTRODUCTION
- Les modes de créativité du langage sont d'une très richesse et d'une telle variété que l'on peut percevoir une modulation propre chez chaque individu, qui, mis à part l'objectivité de l'enonce reflète le sentiment de l'émotivité. On parle alors de l'intonation et de sa fonction expressive, c'est à dire de toute l'information que va au-delà du message référentiel. "Bien souvent pourtant la fonction référentielle a une importance mine et le véritable message peut être décodé que dans la parole proférée" (Leon 1979:159).
- Nous essaierons de montrer, dans cette étude, la fonction phonostylistique de certaines expressions mélodiques de l'intonation dans un groupe de femmes de la société vénézuélienne.

1. CORPUS DE TRAVAIL
- Le corpus utilisé pour cette analyse est un échantillon de la parole de 30 femmes appartenant à différentes classes sociales: favorisée, moyenne et défavorisée, et dans des contextes situationnels différents: participation à des émissions radiophoniques, interviews personnelles, conversation spontanée.
- Dans les échantillons de parole, nous avons sélectionnons un ensemble d'énoncés qui "semblent" caractériser la voix féminine de cette société, dans certaines circonstances.

3. REPRESENTATION DE LA COURBE INTONATIVE
- La mélodie de la phrase est représentée par des graphique comme ceux choisis par Fant (1984) pour indiquer le patron prosodique de la phrase déclarative avec deux, trois et quatre groupes toniques.
- Malgré la grande simplification de ces graphiques, il est possible de représenter les traits intonatifs qui nous intéressent.
- L'analyse intonative de Fant rend compte des oppositions intonatrices initiales, médiales et finales, mais la courbe mélodique se trouve simplifiée par le fait qu'un séquence de syllabes atomiques, quel que soit leur nombre, est toujours représentée comme ayant la même quantité qu'une syllabe tonique. Ce type de représentation met plusieurs phénomènes tels que la pente globale de la courbe mélodique à mesure que l'énoncé se développe.
- Le système de notation de l'ant permet de distinguer quatre niveaux significatifs de la courbe méloïdique: un niveau bas (B) un niveau moyen (M) un niveau haut (H) et un niveau haut extrême (H+).

4. ANALYSE DE L'ÉNONCÉ
- De tout le corpus analysé, nous avons isolé, seulement à titre d'exemple, trois expressions non marquées par le fait phonostylistique que l'on veut faire remarquer, et trois expressions marquées.

4.1. Graphiques de l'intonation "non marquées":
- Les figures (1), (2), (3), présentent l'intonation non marquée.

Les figures (2) et (3), il semble qu'au point le plus élevé ve retrouver sur la tonique sur laquelle on veut insister.

4.2. Graphiques de l'intonation "marquées":
- Les figures (4), (5) et (6) représentent l'intonation "marquée".

- La différence entre ces énoncés marqués apparaît aussi bien à la finale de la courbe mélodique que dans des manifestations de hauteur et de quantité. Le dernier syllabe tonique peut varier d'une manière très faible (fig 4) à une descente considérable (fig 5).
- Sur les graphiques (4), (5) et (6), on peut remarquer un mouvement ascendant-descendant sensible, du au découpage des syllabes avec allongement de la voyelle et en produisant une sorte d'articulation musicale, très perceptible dans le type d'énoncé représenté sur la figure (4). Comme il n'est pas possible de représenter la durée des sons graphiques, il nous faut signaler un allongement sensible de la voyelle en syllabe accentuée, qui s'accompagne, dans la plus part des cas, d'une baisse de l'intensité, telle qu'on peut l'apprécier sur la figure (5).
(toniques 2 et 3) et sur la figure (6) (dernière tonique).
- Les intonations montantes dans les syllabas toniques sont per—
ques comme l'empresse, les courbes
descendantes, comme de la persua—
dation. La diversité de l'effet de
l'intonation est telle que l'as—
pct sémantique de l'expression
est plus important que celui des
mot textuellement représentés.
- Les courbes intonatives mar—
quées et non marquées sont des
fragments de la parole qui ten—
tent de compléter une expression
dans une séquence organisée du
discours global.

5. VARIABLES INTONATIVES ET
FACTORIES PHONOSTYLISTIQUES.
- Des registres correspondant aux
30 locuteurs qui constituent la
totalité de l'échantillon de ce
travail, on a pris 150 phrases
déclaratives de manière à quanti—
fier le pourcentage d'occurrence
de la courbe intonative marquée
et non marquée.
- Le tableau 1 indique les résul—
tats de cette analyse:

<table>
<thead>
<tr>
<th>tableau 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI M 316 27.4 %</td>
</tr>
<tr>
<td>CI non M 834 72.5 %</td>
</tr>
</tbody>
</table>
- les données démontrent que dans
le corpus étudié l'expression inton—
atives marquées est beaucoup
moins fréquente que la non mar—
quée.
- Le tableau 2 indique le nombre
et le pourcentage des occurrences
en fonction de la classe sociale:

<table>
<thead>
<tr>
<th>tableau 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI M 130 (41) 72 (23) 114 (36)</td>
</tr>
<tr>
<td>CI non M 355 (32) 454 (54) 115 (14)</td>
</tr>
</tbody>
</table>
- Comme on le remarquera sur le
tableau 2, le nombre d'occurrentes
de la variable marquées décroît
dans l'ordre suivant: classe favo—
risée, défavorisée, moyenne.
- Il faut maintenant insister sur
le fait que la structure sociale
de cette communauté linguistique
indique que le niveau élevé
qui correspond à un niveau éco—
nomique et non pas intellectuel.
- Le niveau moyen correspond au groupe
intellectuel dans son ensemble.
- La classe défavorisée est celle qui
occure ni au pouvoir écono—
que ni au pouvoir intellectuel.
- Cette structure sociale
permet que l'emploi de certains
patrons intonatifs, comme celui
que nous avons appelé "intona—
tion marquée" serve à eluder ou
à manipuler une situation grâce
de la séduction implicite dans la
mélodie que ce type d'intona—

La lecture de la variable intona—
tives marquées sur la variable
moins fréquente que la non mar—
quée. Dans ce cas, le pourcentage
of la variable intonative
moins fréquente que la non mar—
quée. La classe défavorisée est celle
qui accède au pouvoir écono—
que et au pouvoir intellectuel.
- Cette structure sociale
permet que l'emploi de certains
patrons intonatifs, comme celui
que nous avons appelé "intona—
tion marquée" serve à eluder ou
a manipuler une situation grâce
de la séduction implicite dans la
mélodie que ce type d'intona—

4. On a pur remarquer que
l'emploi de la variable intona—
tives marquées ne correspond pas
exclusivement aux familles d'un
déterminé dans la société
mais que certains groupes
courrent se différencier par
l'emploi plus ou moins prononcé
de cette variable.
- Ce n'est donc pas l'objectivi—	é de la fonction référentielle
dont donne toute l'information,
c'est le message phonostyli—
que qui rend compte du sens
occulte de l'expression et qui
nous permet de conclure que
toute parole est revêtue
d'intention.

7. REFERENCES
dio sociolingüístico de la ento—
nación en el habla de Tacuarembó", Trabajo presentado en el XI ENDI,
San Cristobal.
tiva en español: estudio
sintáctico e intonativo", Acta Uni—
versitatis Upsaliensis.
phonostylistique", Studia Phonéti—
ca, 4, Montréal, Paris, Bruxelles,
Didier.
[4] LEON, P., FAUBE, G., RIGAULT,
A. (édit.), (1970), "Prosodic fea—
ture analysis. Analysis of the
fautes
ETUDE DES PARAMETRES TEMPORELS DES VOIX SANS LARYNX

L. Crevier-Buchanan
U.F.R. Linguistique, Université Paris VII, France.
INSERM, Labo. de Recherche sur le Langage, Paris.

ABSTRACT - Speech timing including voicing events and pauses distribution was evaluated and compared to laryngeal voices. Speakers with tracheo-oesophageal voices using pulmonary air were able to preserve the rhythm and the syntactic/semantic structure of their speeches, whereas speakers with oesophageal voices often needed to insulate the oesophagus and therefore had a staccato-like speech. The phonation time was quite similar in both situations, but the length and the number of the pauses made the difference.

1. INTRODUCTION
Tous les auteurs s’accordent pour dire que les patients utilisant un shunt trachéo-oesophagien (FTO) ont une parole plus agréable que les patients utilisant une voix oesophagienne classique (VO) (1, 4, 5, 6). Nous avons voulu compléter l’analyse des paramètres temporels en étudiant la relation phonation-pauses et la répartition de ces pauses dans le discours.

2. MATERIEL-METHODES
Cette étude a porté sur 12 patients laryngectomisés et 7 témoins de sexe masculin. Parmi les patients, il y avait 2 voix oesophagiennes (VO) (Groupe I), 6 shunt trachéo-oesophagiens autocontinents (FTO) (Groupe II) et 4 prothèses phonatoires (FP) (Groupe III). Deux signaux ont été enregistrés : le signal acoustique et le signal électroglottographique (BGG). Le protocole comprenait des tâches permettant d’explorer différentes situations de parole :
- le temps maximum de phonation (TMP) sur "A", sur une seule expiration ; la durée d’émission d’une phrase (la "phrase") : "C’est une affaire intéressante, qu’en pensez-vous ? Il faut la faire sans aucune regret" ;
- le calcul du nombre de syllabes lu par minute lors de la lecture d’un texte "Grand-mère raconté" (251 syllabes).

On présente ici l’analyse des données temporelles décrivant la tenue de voyelles et de la "phrase" ; pour celle-ci, 3 paramètres ont été retenus :
- la durée totale de la "phrase" ;
- la durée totale de phonation (somme des mots constituant de la "phrase") ;
- la somme des silences entre les mots correspondant à une ponctuation syntaxique et les pauses ne correspondant pas à une telle ponctuation, mais dépassant 160 msec. et se situant entre 2 mots.

La "phrase" était d’une fois par chaque sujet (19 "phrases" analysées) de même pour la voyelle (19 voyelles tenues analysées)

Pour traiter les signaux enregistrés, un équipement informatique Macintosh a été utilisé avec la carte Mac Speech Lab et 2 logiciels "Sound Kit" et "Signalize". Les données statistiques ont été analysées par le programme FCSM traité sur IBM PC compatible. Nous avons traité les variables pour des critères quantitatifs par le test H non paramétrique de Kruskal-Wallis et comparé ces variables 2 à 2 par le test de Mann et Whitney.

3. RESULTATS (Tableau)
3.1. Le temps maximum de phonation
3.1.1. Comparaison entre les patients
Il n’existe pas de différence significative de durée du TMP selon le mode de production du souffle phonatoire (p=0.02).
Les patients du Groupe I avaient un TMP sur une voyelle de plus de 2 sec. En revanche, les patients utilisant de l’air d’origine pulmonaire lors de l’expiration par l’intermédiaire du shunt trachéo-oesophagien (Groupe II et III), avaient des durées d’émission vocale allant de 5 à 11 sec. La différence de TMP n’était pas significative entre les Groupes II et III (p=0.05). Enfin, la différence était statistiquement significative entre le Groupe I et les Groupes II et III.

3.1.2. Comparaison avec les témoins
La différence était significative entre les Groupes I et II et le Groupe témoin, par contre la différence entre le Groupe III et le Groupe témoin n’était pas significative (p=0.08). En d’autres termes, les patients avec une prothèse phonatoire avaient un TMP plus long que le témoin.

3.2. Les variations temporelles dans une situation de parole
3.2.1. Durée totale de la "phrase"
Pour le Groupe I, elle était de 5.37 à 7.78 sec. ; pour les Groupes II et III elle était de 5.33 à 9.89 sec. ; pour le Groupe témoin elle était de 4.2 à 6.1 sec. Les différences de durée de "phrase" entre les 3 Groupes de patients et le Groupe témoin étaient significatives (p=0.006). En effet, lorsque l’on compare les 3 Groupes de laryngectomisés entre eux, ils avaient une durée de phrase équivalente à celle du mode de production.

3.2.2. Durée de phonation
Pour les patients laryngectomisés et le Groupe témoin, une durée de 1er phonation de 27 sec. a été acquise. La durée de la "phrase" s’était légèrement allongée par augmentation de la phase de phonation.

3.3. Le début phonatoire
Il a été calculé à partir du nombre de syllabes lues par minute. On a pu constater que le débit phonatoire des patients laryngectomisés (p=0.006) par rapport aux témoins. Les patients du Groupe III avaient une moyenne plus proche de la normale que les patients du Groupe I et II.

3.2.3. Durée des pauses
Si l’on considère le temps total des pauses et leur répartition, on constate que les différences étaient significatives, tous Groupes confondus (p=0.01).
- Comparaison des patients entre eux : la durée des pauses était allongée dans les 3 Groupes de patients ; les Groupes I et II n’avaient pas de différences significatives entre eux (p=0.02), alors qu’elle était significative avec le Groupe III.
- Comparaison des témoins : il n’exista pas de différence significative entre le Groupe III et le Groupe témoin (p=0.02), alors que la durée des pauses était toujours supérieure à la normale pour les Groupes I et II.

3.4. Étude longitudinale de patients et témoins
Quatre patients ont pu faire l’objet d’un réenregistrement à 6 mois de distance du 1er examen : 2 VO et 2 FTO. Pour les 2 VO et 1 des patients avec une FTO, on a pu faire le contrôle après 12 mois et la durée de la pause phonatoire ; en effet, les temps de pauses nécessaires aux respirations et aux inspirations étaient moins compréhensibles.
L’autre patient avec FTO ne parlait qu’en voix chuchotée lors du 2ème enregistrement. L’intelligibilité était excellente, les variations temporelles comparables à celles de voix laryngées.
Six mois plus tard, lors du 2ème enregistrement, la sonorisation était acquise, la durée de la "phrase" s’était légèrement allongée par augmentation de la phase de phonation.

3.6. Le débit phonatoire
Il a été calculé à partir du nombre de syllabes lues par minute. On a pu constater que le débit phonatoire des patients laryngectomisés (p=0.006) et les patients du Groupe III avaient une moyenne plus proche de la normale que les patients du Groupe I et II.

3.7. Différentes durées de pauses
Pour les témoins, à l’issue de la période d’utilisation des prothèses, une durée de pause phonatoire de 27 sec. a été acquise. La durée de la "phrase" s’était légèrement allongée par augmentation de la phase de phonation.

3.8. Étude longitudinale de patients et témoins
Quatre patients ont pu faire l’objet d’un réenregistrement à 6 mois de distance du 1er examen : 2 VO et 2 FTO. Pour les 2 VO et 1 des patients avec une FTO, on a pu faire le contrôle après 12 mois et la durée de la pause phonatoire ; en effet, les temps de pauses nécessaires aux respirations et aux inspirations étaient moins compréhensibles.
L’autre patient avec FTO ne parlait qu’en voix chuchotée lors du 2ème enregistrement. L’intelligibilité était excellente, les variations temporelles comparables à celles de voix laryngées.
Six mois plus tard, lors du 2ème enregistrement, la sonorisation était acquise, la durée de la "phrase" s’était légèrement allongée par augmentation de la phase de phonation.

3.6. Le débit phonatoire
Il a été calculé à partir du nombre de syllabes lues par minute. On a pu constater que le débit phonatoire des patients laryngectomisés (p=0.006) était plus proche de celle des témoins. Il n’y avait pas de différence significative globale.
4. DISCUSSION
4.1. Le temps maximum de phonation
La durée maximum de phonation reflète les capacités physiologiques d'émission prolongées de voeux.
Il est donc logique que le temps maximum de phonation pour le Groupe I soit bref, car leur volume d'air phonatoire est limité au volume éructé, alors que les patients des Groupes II et III ont une soufflerie proche de la normale (1). La différence de TMP au sein des Groupes II et III peut être expliquée par une fuite d'air lors de l'obturation du trachéostome ou une résistance importante du shunt trachéo-oesophagien au passage de l'air.
De plus, une tension importante du muscle crico-pharyngien peut modifier l'inertie de la néoglottis et l'adaptation de la pression sous néoglottique, responsable de ces variations temporelles.

4.2. En situation de parole
Notre étude a montré en evidence que les patients laryngectomisés élaborent une stratégie de lecture qui se fait aux dépens des temps de pause ; en effet, la durée de phonation n'était pas significativement différente entre les 3 Groupes de patients. On a observé cependant, pour le Groupe I, que le temps de phonation était limité par le volume d'air éructé, les pauses étaient plus nombreuses, correspondant aux raccourcissements et le temps total de pause était allongé. Pour une durée de "phrase" identique pour les 3 Groupes, on constate que le Groupe I avait une durée de phonation raccourcie, les patients prononcent les mots plus rapidement et la somme des pauses est plus importante (2). On pourrait supposer que les patients utilisant la soufflerie pulmonaire, ont une autonomie phonatoire proche de la normale. Les locuteurs prennent le temps de respecter les pauses, de segmenter leur discours selon la structure syntactico-sémantique.
En voix oesophagienne, le discours est scandé, haché par ces interruptions brèves et répétées.

5. CONCLUSION
Deux situations différentes ont été analysées : la durée d'émission d'une voyelle tenue dont les modifications sont physiologiques et les variations temporelles dans une situation de parole, impliquant des stratégies linguistiques ou phonologiques ou morphosyntaxiques.
Le temps maximum de phonation sur une expiration ou une éructation met bien en évidence la différence de mécanisme aéro-dynamique. Le volume d'air éructé est peu modulable.
A l'opposé, l'organisation d'une phrase ou d'un texte dépend de la façon dont le sujet va apprendre à gérer son éructation ou son expiration. Les patients du Groupe I allaient tendance à lire plus vite le mot pour compenser des pauses globalement plus longues ; en fait, il s'agit plutôt de l'augmentation du nombre des pauses courtes lors de chaque injection.
Les patients utilisant la soufflerie pulmonaire ont une autonomie phonatoire proche de la normale (3). Les locuteurs prennent le temps de respecter les pauses, de segmenter leur discours selon la structure syntactico-sémantique. La parole est plus agréable et surtout permet de retrouver les manierismes et les particularités du locuteur (2).

6. REFERENCES

Tableau des résultats

<table>
<thead>
<tr>
<th>Temp. max. (m)</th>
<th>de Phonation (s)</th>
<th>Duree de la Phonation (s)</th>
<th>Duree du Voix (s)</th>
<th>Nombre de Syllabes/Min.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

PHONETIC ASPECTS OF SPEECH PRODUCED WITHOUT A LARYNX

Lennart Nord, Britta Hammarberg* and Elisabet Lundström

Dept of Speech Communication & Music Acoustics, Royal Institute of Technology, KTH, Box 70014, S-10044 Stockholm, Sweden.

Phone 46 8 7907874, Fax 46 8 7907854

*also at Dept of Logopedics and Phoniatrics, Karolinska Institute, Huddinge Hospital, S-141 86 Huddinge, Sweden

ABSTRACT

The aim of the present report is to compare the different types of alaryngeal voices, esophageal and tracheo-esophageal voices, acoustically and perceptually. A general objective is to try to establish the acoustic cues for normal-ness in laryngectomy speech and what constitutes the typical alaryngeal voice quality. Other tasks include intelligibility and acceptability ratings with professionals as well as with naive judges. Analysis of selected aspects are reported, such as voice quality features and prosodic features. Differences and similarities between the voice productions are discussed.

1. INTRODUCTION

After a laryngectomy, the patient has to learn to master speech with a new voice source. The sound generator is the upper part of the esophageal entrance, which is set into vibration by a jet of air that is in-sufflated into the esophagus from the mouth, or taken from the lungs via a tracheo-esophageal fistula. Acoustic and perceptual aspects of the two kinds of speaking techniques, hereafter called "E-speech" and "TE-speech", were compared. Comparisons were also made using characteristics used for descriptions of normal laryngeal speech ("N-speech") [3]. Previous reports have dealt with acoustic and perceptual aspects, see [4-7,10,11].

2. SPEECH MATERIAL AND SPEAKERS

The speech material contained vowels in carrier phrases, sentences with different prosodic patterns, a short informal conversation and a standard Swedish text of 90 words. So far, 6 TE-speakers, 8 E-speakers and 4 normal laryngeal speakers of the same age group (45 - 80 years) have been analyzed. Two of the TE-speakers used Panje voice devices and three low-pressure Blom-Singer devices.

3. PRESSURE AND FLOW MEASUREMENTS

To investigate pressure and flow conditions and also to get an estimate of the voice source shape and spectral content, a flow mask [13] was used in separate readings of /papa/, embedded in a carrier phrase. Subjects were asked to produce these words at three loudness levels, respectively estimated as weak, normal and strong. Inverse filtering and pressure measurements were performed on three E-speakers, three TE-speakers and two normal speakers. Mean values of all /p/ measurements for the three loudness levels and for the three speaker groups were calculated. As can be seen in Figure 1, the normal laryngeal speakers generally produced the words with lower pressure values than what the alaryngeal speakers did, especially when they were asked to produce sounds with low intensity. The alaryngeal speakers could not change their voice levels to the same extent as the laryngeal speakers could, but still managed to vary the loudness level in three steps. Mean values were for the E-speakers 14 cm H2O, for the TE-speakers 22 cm H2O and for the normal speakers 7 cm H2O. This result compares favourably with what is known from investigations of source pressure levels, e.g. [12], in which TE-speakers were found to speak as loudly as laryngeal speakers. E-speakers usually have weaker voices than the others.

Figure 1. Pressure values in cm H2O during the production of /p/ for E-, TE- and normal laryngeal speakers, at three loudness levels, weak, normal and strong. (3 subjects in each speaker group; 15 samples displayed)

4. INVERSE FILTERING AND VOICE QUALITY

By means of inverse-filtering of the air flow during phonation, the periodicity of the wave shapes was analyzed and correlated to perceived voice quality. Flow glottogram curves were obtained for three of the alaryngeal speakers, although they showed a great deal of irregularity. Figure 2, two examples of automatic inverse filter analysis are shown [2].

Figure 2. Flow registrations (upper curve) and corresponding inverse filtered flow glottogram (lower curve) of vowel pulses in /papa/, uttered by a TE-speaker (a) and an E-speaker (b). Inspection of the unfiltered speech oscillogram revealed unusual excitation traces. In Figure 3, vowel excerpts are shown for one E-speaker and one laryngeal voice. As is clearly evident, there is no well-defined single point of excitation for the alaryngeal voice, compared to what is the case for the normal laryngeal voice.

Figure 3. Speech wave oscillograms of vowel samples for an E-speaker and a normal speaker (N).

5. LONG TIME AVERAGE SPECTRA

Long-time-average spectra of these voices have been derived and analyzed. A reading of text passage of approximately 45 secs was used as analysis material. The signal was fed into a Hewlett Packard 3562 A Dynamic Signal Analyzer and spectral analysis was performed. On the spectral display, it was possible to identify the isolated peak corresponding to the level of the fundamental during the reading. We have not discarded the unvoiced segments from the reading, but still consider the result as representative of the spectral distribution and also the relative level of fundamental in comparison with total spectral energy. In Figure 4, LTAS- spectra for a TE-speaker and a normal laryngeal speaker ("N") are shown.

Figure 4. Long Time Average spectra of a read text passage by a TE-speaker and an N-speaker. The level of the fundamental, LD, is indicated by *.

The spectral level difference between fundamental and first formant level (L1-L0), seems to be a valid parameter for these alaryngeal voices. So far, preliminary data from seven alaryngeal speak-
ers, suggest that the L1-L0 difference is larger for the laryngeal voices than for normal voices, i.e. the level of the fundamental is very weak in the laryngeal voices, see Figure 5. Moreover, it does not vary with loudness to the same extent as in normal laryngeal voices [11].

Figure 5. Difference data of total level (Ltotal) to the level of the fundamental (L0) derived from LTAS-analysis of a text passage, read by three E-speakers, four TE-speakers and four N-speakers.

6. PROSODY

Pitch and duration cues

Prosodic studies of intonation patterns and pitch emphasis related to small pitch range and pitch dynamics were made.

In order to evaluate the capability of these speakers to produce acceptable prosodic patterns, a set of sentences with question intonation and emphatic word stress was included in the reading material. In most cases the speakers were able to produce the target sentences. However, they sometimes chose different strategies compared to speakers with laryngeal phonation. Word emphasis was often made by a pausing as well as by a pitch change. In Figure 6 two pitch curves are shown, produced by two laryngeal speakers, one female E-speaker and one male TE-speaker, and an occasional pitch pattern. The problem often is that the aperiodicity creates noise, overlaid on the fundamental. The second pitch curve, displayed in Figure 6. Although varying intensity, the pitch does not exceed 60 Hz (mode value 43 Hz).

7. DISCUSSION

As reported in previous studies on pathological voices, there is a correlation between the voice pulse shape and the perceptual impression of voice quality. An irregular and strongly varying voice source pulse often correlates with a harsh voice [1]. One finding in the present study was the unusual excitation patterns of the laryngeal voices. We still need a better insight into the mechanisms behind these irregular patterns, and a modelling of the structures responsible for these vibrations would be of great value. Work in this area is going on [9].

As a result of the present study so far, two differences between the normal laryngeal and the laryngeal groups are evident. Firstly, the laryngeal voices were characterized by a weaker fundamental relative to the total energy level as compared to the normal voices. Secondly, apart from this static aspect of the voice source, a dynamic aspect is observed if speakers are asked to produce sounds with different intensity. Normal, laryngectomized voices will have a more pronounced fundamental if they phonate at low intensities. The same does not happen for the laryngeal speakers.

8. CONCLUSIONS

It was found that the laryngeal voices, E-voices and TE-voices were characterized by the fundamental compared to normal laryngeal voices. Other, more detailed voice source characteristics, such as inverse filtered flow regenerations displayed strong irregularities for the laryngeal voices.

Acknowledgement: This work was financed by research grants from the Swedish Cancer Society and the Swedish Council for Social Sciences.

REFERENCES

ON USING INTENSITY AS A CODING PARAMETER IN TACTILE SPEECH STIMULI: PSYCHOPHYSIOLOGICAL DISCRIMINABILITY EFFECTS

H. G. Piroth and S. G. Froerer
Institut für Phonetik und Sprachliche Kommunikation der Universität München, Germany

ABSTRACT
In preparation of a system that uses intensity as a coding parameter for tactile speech, this paper reports an investigation of two general psycho-physiological effects that show to be involved in intensity perception, namely the order effect and masking.

1. INTRODUCTION
Not only in psycho-physiology, but also in application-oriented research to establish electrocutaneous speech transmission systems for the deaf, questions concerning the human ability for tactile intensity perception have an important role. In developing an electrocutaneous speech-to-skin communication aid that transmits articulation-based features, we assume that a supra-ergastic component (stress, intonation) could be added to the feature coding method by superimposing intensity variations on the segmental stimuli (9). Classical investigations on electrocutaneous intensity perception discuss the number of possible steps that can be discriminated between absolute threshold and pain. Lindner 1971/1972 has reported that the pain threshold is reached at approximately 5 times absolute threshold. At a frequency of 400 Hz he situates absolute threshold at about 0.2 mA, pain threshold at 1.7 mA with 27 distinguishable steps in between. Schöbel 1936/1937 determined a difference limen of 4 to 5% in normal hearing subjects, Anderson and Munsaw 1951 of 1 to 5% in the frequency range between 100 and 3000 Hz. Hawkes 1959/1963 measured a limen of 3.5% at an intensity of 120 dB and of 3.8% at 2000 above absolute threshold.

A pilot experiment with more complex stimuli in an electrocutaneous pulse train transmission showed that at least two different intensity levels can be identified after a short training period. The present experiment was conducted to gain more knowledge on the discriminability of tactile intensities in complex stimuli. Especially, dependency effects of intensity perception on the temporal and spatial stimulus structure were investigated.

2. APPARATUS
The test stimuli were constructed and presented with the 16-channel System for Electroocutaneous Stimulation SEHR-2. Four rows of electrode pairs were fixed along the dorsal, ulnar, radial, and anterior sides of the Sa's I and X forearm. (See I.I for details and illustrations.)

3. STIMULI
Four complex stimuli were constructed with pulse train sequences as their basic part consisting of three biphasic pulse pairs with a rectangular part in one and a hyperbola-shaped part in the other polarity, resulting in a d.c.- component equaling 0. The pulse repetition rate was 400 Hz. In stimulus 1 eight pulse trains were delivered surrounding the arm at four distal electrode pairs with two succeeding pulse trains at each place and a constant interval of 15 ms after each of the eight pulse trains. The pattern started at the ulnar side and proceeded to the dorsal side. Then, without an intentional pause, a longitudinal sequence of pulse trains was presented oscillating between the distal electrode pair on the dorsal side and the neighbouring dorsal electrode

In this way a 4x2x5 factorial test design was constructed with 4 stimuli, 2 ordinations (ascending and descending intensities) and 5 intensity levels.

One subtest included 10 repetitions of pairs of stimuli I and II (/-/ /-/-/ /-/-/ /-/-/ /-/-/) in randomized order, the other subtest of stimuli III and IV (/-/-/ /-/-/ /-/-/ /-/-/) resulting 400 pairs for each subject. The interval between the pairs was set to 4 s. Eight Ss participated in the experiment. They received both subtests in different sessions with the order of subtests randomized over Ss. Each subtest was presented in two parts of 100 pairs with a short break in between. Ss were informed that the intensity differences were encoded in the "vocalic part" of the stimuli and had to mark the more intensive stimulus of each pair on an answer sheet.

5. RESULTS
Tab. 1 gives the results of a 4x2x5 factorial MANOVA (SPSS: 161/1975) with stimulus, ordering and intensity level as factors. The overall discriminability was 80 ± 5% and the dependency differences were well recognizable. The MANOVA calculation yielded a significant stimulus effect (p<0.005) and a high significant interaction effect of stimulus, order and intensity level and ordering (p<0.001). It can be seen from Fig. 1 that discriminability increases with intensity level for the second ascending pairs (higher intensity in the second stimulus), but decreases with higher intensity level for descending pairs, thus producing the interaction effect. Concerning the main effect of the factor "stimulus" a DUNCAN post-hoc test showed significant differences (p<0.05) between /-/-/ and /-/-/, as well as between /-/-/ and /-/-/, and /-/-/ and /-/-/, as well as between /-/-/ and /-/-/, and /-/-/ and /-/-/ showed a slight (p<0.01) tendency effect (Tab. 2).

Table 1 Results of the Statistical Analysis

<table>
<thead>
<tr>
<th>Factor</th>
<th>d.f.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>STIMULI</td>
<td>19.1</td>
<td>0.04</td>
</tr>
<tr>
<td>ORDERING</td>
<td>19.1</td>
<td>0.09</td>
</tr>
<tr>
<td>LEVEL</td>
<td>19.4</td>
<td>0.07</td>
</tr>
<tr>
<td>ITx ORD</td>
<td>19.1</td>
<td>0.03</td>
</tr>
<tr>
<td>ITx LE</td>
<td>19.4</td>
<td>0.07</td>
</tr>
<tr>
<td>ORD x LE</td>
<td>19.4</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Tab. 2 Results of the Statistical Analysis

<table>
<thead>
<tr>
<th>ITx ORD</th>
<th>ITx LE</th>
<th>ORD x LE</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.1</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td>19.4</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>19.4</td>
<td>0.01</td>
<td>0.04</td>
</tr>
</tbody>
</table>
could be explained under the assumption of forward masking. Since according to the earlier investigations there is a tendency to forward masking and since intensity variations to be discriminated in the present experiment are encoded into the "vocalic" part of the stimuli, intensities of VC-stimuli should be more easily discriminated. The poor recognition if /i:/ could then be explained if /i:/ had a stronger masking effect than /i:/, the CV-stimuli. In explanation the central representations of the stimuli instead of their peripheral characteristics have to be taken into account. Within the framework of this experiment only a first speculative approach to such an explanation can be proposed: The basic unites of the stimuli (pulse trains) were identical in all cases, but they differed in their temporal and spatial relations. Because of the somatotopical representation of body sites the spatial relations should be preserved in building the central representation. But since more distant and more proximal places were stimulated in the "consontantal" (cervical) parts of the stimuli, the conductive velocity of the peripheral fibres may become relevant to determine the central temporal relations. In CV-stimuli the interval between the last pulse train of /i:/ or /j:/ and the first of /i:/ is 15 ms. The distance between the corresponding places of stimulation is 4 cm, but in /i:/ the place changes in proximal, in /j:/ in distal direction when proceeding from the "consontantal" to the "vocalic" part. Relying on the values given in the literature (13,33) conduction velocity in thick myelinated fibres is between 40 and more than 100 m/s, i.e. even with 40 m/s a distance of 4 cm in the distal-proximal direction produces a change of the temporal intervals of only 1 ms which is too small to cause an effect as observed. But if - as can be supposed - a part of the central representation of the stimulus is based on information processed via thin unmyelinated C-fibres with a conduction velocity of more than 2.55 m/s the temporal intervals at the points of central occurrence of two successive pulse trains at places 4 cm apart from one another differ from the peripheral interval by at least 15.7 ms. Thus, in /j:/ with

\[ /i:/ \text{ being presented at more proximal places the inter-pulse-train interval is centrally doubled (15 ms = 15.7 ms = 30.7 ms), and if /i:/ starting at the distal places it is reduced to approximately 0 (15 ms - 15.7 ms = -0.7 ms). Based on this speculative assumption one could /i:/ and /j:/ cannot cause forward masking, since the vocalic part is presented first (81.00X and 83.88X correct discrimination), (i) /i:/ and /j:/ produces a forward masking effect, since the central representation of /i:/ is built up before the representation of /j:/ is started (thus, only 75.75% correct answers), (ii) For /i:/, the representation of both parts are not separate, but as the central point of occurrence of the last pulse train of /i:/ is identical with that of the first in /j:/ the whole stimulus elicits a unique, more complex representation which is not affected by forward masking (81.49% correct answers). To summarize, the stimulus effect can be explained in terms of central temporal characteristics if C-fibre conduction contributes to the representation of the stimuli used and if forward, but not simultaneous masking is involved in a perceptual process that separates the longitudinal and circumferent parts of the patterns. To evaluate this proposal, more specific electro- or psychophysical experiments are mandatory.

7. REFERENCES


---

**Figure 1:** Discriminability dependent on intensity levels (full squares: ascending, open squares: descending).

---

**Table 2:** Discrimination Dependence on Stimuli

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>/i:/</th>
<th>/j:/</th>
<th>/i:/</th>
<th>/j:/</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>83.88</td>
<td>81.49</td>
<td>81.00</td>
<td>75.75</td>
</tr>
<tr>
<td>/i:/</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>/j:/</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
</tbody>
</table>

---

**6. DISCUSSION**

The interaction of intensity level and ordering shows an order effect as it is known from classical investigations on the perception of temporal durations (e.g. T2). In general, this so-called time-order error produces discrimination rates that are dependent on the order of the stimuli and on the duration of the inter-stimulus interval between them. A similar effect in the discrimination of the durations of tactile stimuli was found by Pirotta-Tillmann (1987) [8], thus it is clear now that duration as well as intensity perception of electro-tactile stimuli is affected by time-order error.

The asymmetric dependency of discrimination rate on the kind of stimulus presented, is more difficult to explain, namely the significantly low results for stimulus /i:/ /j:/. The rank order of the stimuli shows that intensity discrimination tends to be better with /i:/ than with /j:/ and better with VC than with CV. Similarly, Pirotta 1986 (7) had shown that identification of tactile vowels is higher in VC and identification of consonants is higher in CV syllables. Those effects
SPÉECH PERCEPTION ABILITIES OF PATIENTS USING COCHLEAR IMPLANTS, VIBROTACTILE AIDS AND HEARING AIDS

Eva Agefors and Arne Råsberg
Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Box 70014, SE-100 44 Stockholm, Sweden

ABSTRACT
The speech perception ability reported from profound hearing impaired persons using different technical aids: hearing aids, cochlear implants or tactile aids, varies widely. A test-battery was constructed that consisted of segmental and suprasegmental tasks and speech tracking. Two presentation modalities were used, vision-only and visual information supplemented with the assistive device. Three groups of subjects participated, deafened adults, subjects with profound postlingual hearing loss and normally hearing subjects artificially deafened. The results indicated that use of a hearing aid by listeners with some residual hearing provided more information than the other assistive devices.

1. INTRODUCTION
During the last two decades the research in the fields of electronics, audiology, speech science and surgery has made it possible to introduce a limited world of sound to many profoundly hearing impaired and deaf persons. This has been carried out by more sophisticated and powerful hearing aids or by cochlear implants, which directly stimulate the auditory nerve or by tactile aids, which employ the cutaneous sense and its pathways for transferring information. The aim of this study was to develop a simple test battery and to compare the effectiveness of tactile aids, hearing aids and cochlear implants. It is recognized that the comparison between results obtained by different teams or devices in tests with the postlingually deaf is difficult. To get an uniform selection of patients it was more or less impossible. The performance among individuals shows often great variations, not only as a result of what they hear or feel with their device, but also as a result of their varying ability to lip-read or make use of small linguistic and paralinguistic cues. A standardized test battery does not exist, so any language and the phonological characteristics from one language to another make the interlinguual comparisons complicated.

During the last years, research groups have reported that prosodic features, such as syllable length, stress pattern and vowel length, as well as segmental features such as voicing and manner of articulation may be transmitted through the tactile modality [6]. A few studies have also reported good tactile support during speechreading of normal speech [8].

Great variations among patients using the same type of cochlear implant have been reported, but results from both single-channel users and multichannel users show that the devices can provide important cues to intonation, manner and voicing that are significant to lip-reading [1].

In some patients very good speech understanding with or without support of lip-reading has been reported from cochlear implanted patients using either single-channel [7] devices or multi-channel devices. Dowell et al. [3] have reported that 50% of the patients using (Nucleus) multichannel cochlear implants have demonstrated ability to understand connected discourse with auditory input only.

2. SUBJECTS, MATERIALS AND METHODS
Four different groups of subjects participated voluntarily in the testing. In the vibrotactile group eight subjects participated (Vt1-1:Vt8). Three deafened adults (Vt1-1:Vt3) had varying experience of tactile aids. Five normally hearing subjects were artificially deafened and had experience of about 100 hrs of training with vibrotactile aids. Two vibrotactile single-channel aids were used, an ordinary bone-conductor coupled to an amplifier (6 subjects) and the Minivib (2 subjects). The processor in the Minivib gives amplitude modulated pulses at a fixed frequency of 220 Hz. The acoustic energy at the frequencies between 700 and 1500 Hz is extracted. During testing the subjects held the vibrators in their left hand.

In the cochlear-implanted group, six subjects participated (Ci1-Ci6). Two subjects, Ci1 and Ci2, were implanted with a single-channel extra-cochlear implant (Wien/3M) and four subjects were implanted with a multichannel intra-cochlear implant (Nucleus). Subjects ranged in age from 36-65 years and they represented an average sample of adults, who had received cochlear implants in Sweden. The cochlear implant users had a daily experience of their devices from 6 months up to 5 years.

In the hearing aid users group, eleven subjects participated (H1:H14) and (H15:H17). Subjects ranged in age from 38-75 years and they were all profoundly hearing-impaired since many years. During testing they wore their own hearing aids. Although all subjects were profoundly impaired, the subjects were not equivalent audiometrically. For that reason they were divided into two groups: group H1 had mean hearing loss at frequencies 500, 1000 and 2000 Hz of 104 dB, sd 13.1 dB and group H2 with mean hearing losses of 82 dBm, sd 16.1 dB.

In the normally hearing group four subjects with simulated hearing-loss participated (Lpl:LP4). They listened to low-pass filtered speech at cutoff frequencies 250, 5 and 1 kHz. The filter had a damping of more than 80 dB/oct. White noise was added, S/N = 20 dB. The subjects ranged in age from 25-45 years.

The test material consisted of three parts: Intervocalic consonants, prosodic contrasts and repetition tracking. The segmental test used a set of 16 CV utterances with a carrier phrase in which the vowel was always /a/. Consonants were chosen to sample a variety of distinctions in voicing, place of articulation and manner of articulation.

The suprasegmental test used was a closed-set test battery, presented as a two alternative forced-choice task. The specific prosodic features tested were: number of syllables, voice—length, juncture, tone and emphasis.

Speech tracking was introduced by De Filippo and Scott [4] and has been used to train and evaluate the reception of connected speech via lip-reading combined with different assistive devices. The speaker reads, at a normal rate, sentence by sentence from a book, and the speech-reader (the subject) is required to repeat the information verbatim. If the sentence is not correctly repeated, the speaker employs a hierarchy of strategies to assist the subject in repeating every word correctly. The speech material used was taken from a book by a famous Swedish author. This material was chosen because it has a relatively consistent level of reading difficulty from sentence to sentence. During each test session, tracking was performed for a total of ten minutes under each of two conditions: (a) lip-reading plus aid and alone. The result of the test in words per minute (wpm) was calculated by dividing the number of words correctly repeated by 10 for each ten-minute tracking period. The tracking rate achieved by normally hearing subjects (unmasked) using the same method with the same speaker and the same test text was 88 wpm. The consonant and prosodic tests were videotaped and the speech tracking was presented live. The same speaker, a woman, was used in all test sessions. Each subject was tested individually.

The test order was the same for all subjects: CV syllables, prosody and speech tracking. Each test started with the combined situation.

The normally hearing subjects (Vt group) were masked by earplugs and a noise in the room was used to simulate the auditory environment and aid during the speech tracking situation they were sitting in a sound attenuating testing booth behind and the speaker through a window. The cochlear-implanted subjects and the hearing aided subjects were tested in free...
field at the most comfortable level, adjusted by themselves, in condition lip-reading plus aid. In the situation lip-reading alone the hearing aided subjects were unaided and sitting in the test room under the same condition as the normally hearing subjects.

3. RESULTS AND DISCUSSION

Confusion matrices were constructed for each individual and for each condition. An information transfer measure [5] was calculated for each feature. Three major articulatory and phonetic categories were used: manner (stop, frication and nasality), place and voicing. The results obtained from the segmental test, expressed as mean percent transmitted information of vCr-syllables displayed for each group of subjects in the two conditions are shown in figure 1.

The cochlear implant group was helped by transmitted information concerning the features tone and juncture. These features are among the most difficult to lip-read.

Results obtained from speech tracking are shown in figure 2. The enhancement of lip-reading with the single-channel vibrotactile aid is close to 5 wpm, and about 10 wpm for the group H1. The mean score enhancement for the cochlear implant users was about 25 wpm and about 55 wpm for the group H2. The speech tracking score for the Lp-1000 group reaches the ceiling rate in this particular situation. Data obtained with the speech tracking procedure, clearly show the difference between communication with the vibrotactile aid, cochlear implant and hearing aids.

4. CONCLUSION

The results in fig. 1 and 2 show that the hearing aid using group with a profound loss get very little benefit from their hearing. They might therefore be considered as candidates for a cochlear implant operation. On the other hand, the results also show a large variation in results on all tests for the cochlear implant group. By the use of diagnostic tests of the type presented here, it might be possible to understand the reason for these variation. The results can also be used in patient selection for implantation.

5. ACKNOWLEDGEMENTS

This project has been supported in part by grants from The Swedish Board for Technical Development, (STU).

6. REFERENCES

ABSTRACT

Three postlingually deafened adults who received cochlear implants read passages before and after their prostheses were activated. Their lung volumes were measured with an inductive plethysmograph that transduced the cross-sectional areas of the chest and abdomen. The activation of the cochlear prostheses was followed in every case by a significant change in average airflow, which rose for two subjects with initially low flow rates and fell for one subject with a higher flow rate pre-implant [1].

1. INTRODUCTION

We have been studying speech breathing in late deafened adults as part of a larger project in which we examine physiological and acoustic properties of their speech while they perform a variety of speech tasks, before and after receiving electrical stimulation of the auditory nerve from a cochlear prosthesis. All three subjects became totally deaf in their twenties or thirties with profound bilateral sensorineural losses. Pre-implant they performed at chance levels on auditory tests of closed-set word recognition. Post-implant all three subjects improved in word and sentence recognition.

2. PROCEDURE

In each session the subject read the elicitation passage three times at 20-minute intervals. Subjects F1 and F2 read the Rainbow Passage; M1 read "A Trip to the Zoo". There were two pre-stimulation baseline recording sessions. Then the subjects began to receive electrical stimulation from their Intraid multichannel cochlear implants, and additional recordings were made at intervals of approximately 1, 4, 12, and 24 weeks post-stimulation. The subjects did not receive auditory training or speech therapy. To obtain volumetric measures of speech breathing, we measured changes in the cross-sectional area of the rib cage and abdomen with an inductive plethysmograph (Respiracal). To compute the change in lung volume resulting from a respiratory maneuver, the two amplifier outputs from the plethysmograph are summed after weighting by correction factors. To determine the correct proportion of the two signals for a given recording session, the subject had to perform isovolume maneuvers at the beginning and again at the end of each session. To arrive at a scale factor for converting the summed volume signal to milliliters, the subject exhaled and inhaled into a plastic bag of calibrated volume. Amplified signals from the Respiracal and the microphone were recorded and low-pass filtered and digitized simultaneously. An operator labeled the beginning and end points of expiratory limbs while listening to the synchronized acoustic signal. The labeled events were automatically written into a file which was later accessed for calculating limb duration and limb initiation and termination levels in milliliters re FRC (tidal end respiration level).

3. RESULTS

Figure 1 presents means of average airflow (left column) and volume of air expended per syllable obtained in two sessions prior to receiving stimulation...
Figure 2. Average airflow and volume of air expended per syllable before and after cochlear prosthesis with postlingually deafened adults.

from cochlear prostheses and in four (M1,F2) or six (F1) sessions subsequently. M1, hearing-impaired since birth, averaged initially 181 mL/sec of airflow. After two weeks stimulation from his prosthesis (onset indicated by vertical line) M1 had reduced his average airflow by 15% (third session). On the average, M1's flow rates, after his processor was turned on, were 17% lower than before stimulation (F(1,2) = 22.6, p < .05). In sessions 1 and 2, M1 expended an average of 68.2 mL/syl. After activation of the processor, the volume of air expended fell on the average over four sessions to 46 mL/syl, a decrement of 33% (F(1,2) = 242.6, p < .01). Prior to implant, M1 ended his expiratory limbs 79.6 mL below FRC. This is characteristic of congenitally deaf speakers. Two weeks after activation of the processor, M1's termination level fluctuated around FRC (F(1,2) = 23.0, p < .05). It appears that M1 used his newfound economy of average airflow when reading following implant mostly to desist drawing on expiratory reserve volume (the volume below FRC).

Subject F1, a female, initially expended air during reading at abnormally low rates, averaging 92.0 mL/sec. Following the onset of stimulation, her average airflow increased gradually and irregularly, attaining 144.0 mL/sec after 85 weeks. The mean airflow in all recordings following activation of the processor was 104.3 mL/sec, an increase of 13.4% over the two baseline sessions (F(1,2) = 23.9, p < .05). We observed informally that F1's voice quality has also changed: before stimulation, it was harsh and loud; now it is much softer. The volume of air that F1 expended per syllable increased following activation of the processor by 7.9%, from 23.9 to 25.8 mL (F(1,2) = 33.2, p < .05).

Subject F2 also started out with abnormally low rates of average airflow while reading. Following stimulation, her average airflow increased 20.2%, from an average of 125.8 mL/sec for the two baseline sessions to 151.2 mL/sec for recordings pooled over the four sessions post-implant (F(1,2) = 537.4, p < .01).

4. DISCUSSION

Figure 2 plots mean average airflow (left) and volume per syllable expended before and after activation of the implant. Insomuch as our three speakers are representative of postlingually deafened adults, it appears that the effects of total sensorineural hearing loss in adulthood include anomalies in the management of speech breathing and that these may involve either an expenditure of too much air or too little. Once the speakers received some auditory input from their cochlear prostheses, in every case they modified their speech breathing in the direction of normality. Significant changes were observed in average airflow (M1,F1,F2), in volume of air expended per syllable (M1,F1), and in speech termination levels (M1). Some of the changes in the acoustic correlates of speech that are associated with sudden hearing loss may be mediated by abnormal patterns of speech respiration and laryngeal control of the breath stream. Similarly, some of the acoustic changes that take place when partial self-hearing is restored by cochlear prosthesis may be mediated by a normalization of breath stream mechanisms such as observed in this study. Improper laryngeal valving is a prime suspect in the search for mechanisms underlying the excessive air expenditure of some late deafened speakers and the inadequate air volumes and flow rates of others.

REFERENCE

COMPENSATORY ARTICULATION AND NASAL EMISSION OF AIR IN CLEFT PALATE DISORDERS: WITH SPECIAL REFERENCE TO THE REINFORCEMENT THEORY

Birgit Rutters and Kirsten Brøndsted

Institute of General and Applied Linguistics
University of Copenhagen and
Copenhagen Institute for Speech Disorders

ABSTRACT

It has been assumed that the compensatory speech habits developed in some children born with cleft palate may to some extent be explained by a reinforcement effect induced by the environment on the speech of the cleft palate child due to a perceptual preference of its environment. This is called the reinforcement theory. The results of the present study seem to support the theory and that the mother's education may be one relevant factor.

1. INTRODUCTION

Speech produced by speakers with velopharyngeal insufficiency is always more or less characterized by nasalization. Further, the speech is frequently characterized by nasal emission of air influencing primarily the obstruents ('pressure consonants'). However, some children develop compensatory sounds in the sense that obstruents normally produced at or in front of the velopharyngeal valve are here produced behind this valve. The resulting compensatory sounds are primarily glottal stops and pharyngeal fricatives. The way of speaking results in more or less unintelligible speech. On the other hand, on the surface it may seem more distinct to the listener than speech dominated by nasal emission of air.

It has been hypothesized that compensatory speech is almost always learned and reinforced in infancy and early childhood (1). In other words, according to this assumption the compensatory speech habits learned during language development may be due to the perceptual preference of its environment. The theory to the effect that perceptual preference of the environment may be a reinforcement effect on the speech of the cleft palate child is called the reinforcement theory.

The purpose of the present study is to investigate the parental perceptual preference between compensatory articulation and nasal emission of air in order to deliver support for or to invalidate the reinforcement theory. Since only some (few) children born with cleft palate, rather than most of them, develop compensatory articulation patterns, listeners are, according to the theory, supposed to differ as to their preference of cleft palate speech mode. Thus, it seems relevant to determine if there are some factors which correlate with the parental perception. One such factor could be the social status of the parents, another the parents' education. Further, in order to be comparable with parents of new-born children with cleft palate, the listeners should be parents of normal children, since both of these groups are supposed to be equally unfamiliar with cleft palate speech.

The compensatory articulation starts and progresses during the babbling period and in the very early speech period, where intelligibility in a narrow linguistic sense is irrelevant. Therefore, in order to eliminate the influence from the different intelligibility of the two speech modes, the parents were asked to listen to nonsense words.

2. METHODS

The test included 10 different nonsense words said in the two speech modes. Both speech modes were clearly hypernasal and the most frequent compensatory sound was a glottal stop. The parent listeners comprised only mothers as the mother is normally more in contact with the baby than the father, and thus has greater influence on the child's development, including its linguistic development. The listeners were distributed as follows: 31 mothers; 10 non-educated female cleaners and 11 female secretaries. The mothers were categorized into three groups according to income and into three groups according to education. The teachers and cleaners were included in order to highlight the education factor.

The test tape was individually presented to each subject and the question was: 'Which of the two pronunciation would you prefer if you were talking with the speaker that you hear on the tape?'.

3. RESULTS

In the following, C and E are used for 'compensatory articulation' and 'nasal emission of air', respectively. The results of the C-answers in per cent of the total numbers are depicted in the figure. In general the listeners prefer the E- pronunciation as the C-score is less than 50% averaged over all the listeners, but differences among various groups of listeners may be observed. With the group of mothers there are 38% C-answers, but a clear intergroup variation is seen: In the highest income group there are 59% C-answers, in the average income group 32%, and in the lowest income group 20%. Thus, the number of C-answers given by the mothers seems to be somehow related to their social status, even though only the difference between the high and the low income is clear-significant. Also, the behaviour of the mothers varies according to their education: (1) educational/social training, (2) university training, and (3) other. It is seen that the C-score is highest with the mothers with educational/social training (56%), followed by the mothers with university training (36%), and the group (28%), even though only the groups with the highest and the lowest scores are significant. Finally, the C-score is substantially higher with the teachers.
than with the cleaners, and the difference is highly significant. Notice that the behaviour of the teachers are evidently different from all other groups, whereas the behaviour of the cleaners is within the range of the mothers.

4. DISCUSSION
The purpose of the present study was to throw light on the following question: Is the parental preference between compensatory articulation and nasal emission of air influenced by social status and education? From the results it can safely be concluded that the mothers do not behave alike in their choice between the two speech modes, and that one factor seems to be the social status of the listeners, at least when defined as level of income. Also, the results seem to indicate that education may be a relevant factor. It should be noticed that the mothers with university training and the group including other types of training tend to behave very much alike. This indicates that it is the specific type of training that is the relevant factor, rather than the level of training, even though the few data in the group of university mothers should be taken into consideration. But the finding that the scores obtained by the non-educated cleaners is very similar to the scores obtained with these two groups of mothers, also points in the same direction. It should be added that there is no simple relationship between the three categories of social status and the three categories of education.

Now, do the results support the reinforcement theory? Three groups were more inclined to choose the compensatory speech mode, namely mothers of high social status, mothers with educational/social training, and school teachers. Thus, mothers belonging to these groups should be potential candidates for reinforcing speech with compensatory articulation. Therefore, we checked the files covering a period of 25 years regarding the distribution of cleft palate children with and without glottal stop compensations on mothers of high versus low social status and mothers with educational/social training versus other kinds of training. As to the educational factor, the occurrence of glottal stop compensations are significantly higher with the children of educationally/socially trained mothers than with the other group including children of mothers with university training and other trainings. On the contrary, the mothers in this group only slightly higher occurrence of glottal stop compensations with the high than with the low income group, and the difference is not significant.

Finally, some American studies (1, 3) apparently also deal with parental preference and the two cleft palate types. But after we have listened to the American test tape we think they have examined other speech phenomena. This stresses the need for international agreement on definition of universal speech symptoms, so that research can be compared.

To conclude, the results of the present study seem to support the assumption that reinforcement may be a relevant factor and that the type of mother's education may be a reinforcing element. But it should be emphasized that the causal relation between the two kinds of observations - preference and frequency of occurrence within specific groups - is not necessarily one of reinforcement. It is probably too simplistic to assume that reinforcement, if it occurs at all, is the singular, or even the strongest, factor influencing the development of compensatory articulation. But apart from the conclusions about the reinforcement factor, which is drawn from the current study of preference, it is interesting that listeners' preference between the two deviant speech modes differs according to education and social status. It has been shown that listeners' judgments about the speakers personality and appearance are more negative when listening to voice disorders, including hypernasality, than to normal voice quality. Therefore, it seems likely that when unaware of the poor intelligibility of compensatory speech some listeners may find it more positive (or less negative) than speech with nasal emission of air. But as far as the current study does not report on the relationship between such judgments and the social status and the education of the listeners.

5. REFERENCES
PERCEPTUAL AND ACOUSTIC ANALYSIS OF THE VOICE IN ACUTE LARYNGITIS

Anders Löfqvist and Lucyna Schalen

Department of Logopedics and Phoniatrics, Lund University, Lund, Sweden

ABSTRACT

Acoustic and perceptual analyses of the voice were made in 20 cases of laryngitis during the acute stage and after recovery. The acoustic analysis indicated considerable variability in the pattern of change between the acute and the control condition. However, the perceptual analysis showed consistent, and significant, differences in the ratings of the two conditions. Correlations between acoustic measures and perceptual ratings were generally low.

1. INTRODUCTION

One goal of applied voice analysis is the development of acoustic measurements that are useful for the clinical management of voice disorders. Such measures may supplement current perceptual evaluations used in clinical analysis of dysphonia. Their main advantage is that results obtained at different places can be compared. This is not necessarily the case for the results of perceptual analysis, where the background and training of the listeners influence the results [1-2]. The present study compares perceptual evaluations and acoustic measurements of dysphonia in acute laryngitis.

2. METHODS

2.1 Material

Voice samples from 20 adults (11 females and 9 males) with dysphonia due to acute infectious laryngitis were analyzed.

2.2 Procedure

Voice recordings were made under standardized conditions in a sound-proof room during the acute stage, and a control recording at least two months later. A short story served as the speech material. The duration of the recorded speech was approximately 40 s.

The perceptual evaluation was made by a group of four experienced clinicians using a 5-point rating scale, where 0 represented normal, and 4 maximal deviation. The evaluation comprised 12 different voice qualities. Of these, only those were used in the present study that met two criteria: a significant test-retest correlation, and a significant interjudge reliability (Kendall W). The qualities used here were: dysphonia, breathiness, roughness, aphonia, and voice breaks. In addition, vocal fry was also included, although it failed to show a significant test-retest correlation.

Two different acoustical analyses were made. First, long-time-average spectra were calculated using the procedure described in [7]. This analysis was made of the whole recording, excluding pauses and voiceless segments of the speech signal. Based on this analysis, a rough measure of the tilt of the source spectrum was obtained by the ratio of energy in the frequency bands 0-1 and 1-5 kHz. In addition, the relative energy level in the frequency range 5-8 kHz was calculated; this measure is related to the presence or absence of noise in the voice [9]. Second, the relationship between non-harmonic to harmonic energy (N/S) was estimated using the procedure described in [8]. Due to the computational complexity of this procedure, this analysis only covered a single stressed vowel in the recording; its duration was in the range 106-150 ms.

3. RESULTS

The results of the perceptual analysis are shown in Figure 1. For all six voice qualities, there were significant differences between voice in the acute and the control conditions. All but one of the qualities showed a decrease from the acute to the control stage; the exception was vocal fry.

While the perceptual analysis indicated that there were significant group differences between the acute and the control conditions, the results of the acoustic analysis showed non-significant group differences between the two conditions. The reason is that different voices showed different acoustic patterns of change between the acute and the control condition. This is illustrated in Figures 2 and 3. Here, the voices have been divided into two groups based on the pattern of change revealed by the long-time spectral analysis. Thus, the top part of Figure 2 shows 9 voices where the predominant change is a decrease in the relative energy between 5-8 kHz. The difference between conditions is significant, t(16) = 4.123, p<0.05.

The lower part of Figure 2 shows the remaining 11 voices, where the major change is a decrease in the ratio of energy 0-1/1-5 kHz; also this change is significant, t(10) = 4.539, p<0.01.

Similar results were found for the relationship between harmonic and nonharmonic components in the voice. The top and lower panels of Figure 3 plots the results of N/S for the acute and control conditions for two groups of voices. These groups correspond to the ones shown in the top and lower part of Figure 2, respectively. As shown in the top panel of Figure 3, 8 voices in this group showed a decrease in the N/S from the acute to the control condition. The difference between conditions is significant, t(4) = 2.168, p<0.05. For the remaining 11 voices, the lower panel of Figure 3 shows an increase of N/S from the acute to the control conditions for 8 of them; the difference is not significant, however.

Pearson product-moment correlations were calculated between the acoustical measures and the perceptual ratings. Significant correlations were found between the rating of breathiness and the relative energy level between 5-8 kHz (r = .43, p<0.01), vocal fry and the relative energy level between 5-8 kHz (r = .38, p<0.05), and roughness and N/S (r = .5, p<0.01). The correlations between the
The acoustic measures we have applied are related to the frequency domain. The perceptual qualities of voice breaks, apnea, and diplopia are most likely related to temporal properties of the voice source. Hence, we should note, furthermore, that the acoustic measures we have applied are likely to be highly correlated with the present set of acoustic measures. In addition, the psychophysics of voice evaluation is far from understood, given the complexity of the signal.

Some studies have shown quite significant correlations between acoustic measurements and perceptual ratings [3, 4, 5, 6]. However, the highest correlations are usually found between acoustic measurements and perceptual "supercategories", based on factor analysis or composite measures. When simple perceptual qualities are used, as in the present study, correlations tend to be reduced.

5. ACKNOWLEDGMENTS
This work was supported by funds from the Faculty of Medicine, Lund University.

6. REFERENCES
THE DEVELOPMENT OF ARTICULATORY SKILLS IN CLEFT PALATE BABIES

Kino Janssen-Schultheiss

Dept. of Phonetics, Academic Medical Center,
Meibergdreef 9, 1105 AZ Amsterdam-Zuidoost, The Netherlands
and Institute of Phonetic Sciences, University of Amsterdam.
Herengracht 338, 1016 CG Amsterdam, The Netherlands

ABSTRACT
A description is given of the speech motor (articulatory) development in 3 cleft palate and 2 normal born infants in the first two years of life. The impact of an articulatory impairment of the child speech upon the verbal reactions by the mother is also discussed.

1. INTRODUCTION
In the research project The Influence of an Oral Plate upon Speech Development and Interaction in the First Years of Life of Cleft Palate (CP) Babies' 12 infants, with a complete or isolated cleft palate and 6 normal born babies were studied monthly while interacting with their mothers in a naturalistic, free play situation. Their communicative development (from 0-2:0 years of age) was recorded on video recordings of 20 minutes each. Besides that, a larger group of 40 2:0 toddlers (30 CP and 10 normal born, including the longitudinal group) was recorded once [1]. It turned out that, within this 2 year old group, the CP children without an oral plate (17) uttered less meaningful words, had a less high M.L.U. (L.) (i.e. mean length of (longest) utterance(s), as measured in morphemes), and were less advanced in the use of specific phonological processes than the CP children with an oral plate (13). In comparison with their normal peers, the CP children established far less phonetic, phonological, and syntactic abilities. Looking at interaction, the mothers of the normal born children facilitated the learning process concerning the articulatory proficiency far more by verbal modelling and imitations than the mothers in the CP group. However, the normal born children imitated less than the CP group of children. In our opinion, 'understandability' of the child endeavoured the speech learning process in the child. In the present study the question was raised whether the quality of articulatory development, in terms of speech motor milestones and certain distinctive features, had an impact upon the point of time that the so-called word border (10 or more varied words within five minutes speech sample) was reached, as well upon specific strategies in the mother to reinforce specific articulations of the child by imitating or other verbal reaction upon child speech.

2. PROCEDURES

2.1. Subjects
The speech of 5 children (3 CP and 2 normal born infants) in interaction with the mother has been studied so far.

2.2. Transcription
From each twenty minutes speech registration those five minutes in which the child produced most utterances, were selected. The speech of mother and child was transcribed according to specific codes [1]. In that system the infant speech productions are seen as an oral physiological development with specific stages and milestones. Those go first from laryngeal to single articulatory speech movements and from babbling to the first words. In the case of articulatory movements, the speech output was transcribed in terms of [+ anterior, + explosive and + fricative]. As 'meaningful' word we considered first and all those articulatory strings on which the mother responded by imitating or giving an associated verbal reaction. Furthermore, when the trained transcribers heard a word, either based upon their knowledge of Cleft Palate speech or interpreted from the video picture.

3. RESULTS

3.1. Speech motor aspects
Looking at the overall picture of speech movements in development over the whole period of two years, the CP children differ remarkably from the normal ones (see Table 1). They produce far more laryngeal than articulatory movements. The expression of words did not seem to be related to the amount of articulatory productions in the first two years of age.

Table 1. Overview in percentages (%) of laryngeal (la) and all articulatory sounds (ar) including babbling as well as words (w) and imitations (i) within the first two years of life of 3 CP (+Ch) and 2 normal born (-Ch) children (Ch).

<table>
<thead>
<tr>
<th></th>
<th>Ch</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>la</td>
<td>0</td>
<td>60</td>
<td>53</td>
<td>56</td>
<td>25</td>
</tr>
<tr>
<td>ar</td>
<td>4</td>
<td>12</td>
<td>26</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>w</td>
<td>0</td>
<td>20</td>
<td>15</td>
<td>19</td>
<td>44</td>
</tr>
<tr>
<td>i</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>12</td>
</tr>
</tbody>
</table>

3.2. Articulatory aspects
As shown in Table 2, the CP children have less anterior and more posterior single articulations. Concerning babbling there is variation in general.

Table 2. Overview in % of single articulatory (a) and babbling (b) speech movements (anterior, posterior and varied) as well as words (w), in 3 CP (Ch) and 2 normal children (Ch) measured in the period of 0.2 until 2.0 years of age.

<table>
<thead>
<tr>
<th></th>
<th>Ch</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>w</td>
<td>0</td>
<td>42</td>
<td>41</td>
<td>35</td>
<td>28</td>
</tr>
<tr>
<td>i</td>
<td>4</td>
<td>15</td>
<td>14</td>
<td>41</td>
<td>20</td>
</tr>
<tr>
<td>r</td>
<td>1</td>
<td>41</td>
<td>14</td>
<td>41</td>
<td>24</td>
</tr>
</tbody>
</table>

In Table 3. all the single articulation movements (tokens), also in babbling, were counted and classified in types as well as specific features. We focused upon the anterior articulations, especially the plosives and fricatives. The normal born children produced not only more articulatory movements in general, they produced also a larger variation in articulation types, compared with the articulatory production of the CP child analyzed so far (Table 3). The normal born children produced more anterior articulations in absolute frequency as well as percentages than in one of the CP children, analyzed so far. The speech sounds with the features [+ anterior, + fricative or + plosive], have a high frequency in Dutch and should have - in our opinion - an impact upon the expression of the first words, the point of time in which the word border is reached (see also Table 5).

Table 3. Overview of the total amount of articulations (Na), the number of different articulation types (Nat), and the number of plosives (Nap) as well as fricatives (Nat). Designed as Frequent (Naf), less frequent (Nat), and very frequent (Nap). The ratio of anterior plosives and fricatives with other articulations (% in 1 CP (Ch) and 2 normal children (Ch) from week 10-77).

<table>
<thead>
<tr>
<th></th>
<th>Ch</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Na</td>
<td>136</td>
<td>591</td>
<td>1054</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nat</td>
<td>14</td>
<td>38</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nap</td>
<td>2</td>
<td>193</td>
<td>435</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naf</td>
<td>3</td>
<td>14</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% ant.</td>
<td>2</td>
<td>35</td>
<td>43</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Looking at the interaction between mother and child, we wondered how the mother would strengthen the correct articulations by verbal reinforcement of child articulations (Table 4.), which strategy she would use. At this moment only the material of the two normal born children has been analyzed.

Table 4. Overview of maternal reinforcement of child articulations in absolute frequency, total percentages of reinforced articulations (Na and percentages (%): the number of verbally modelled articulation types (Nat), anterior plosives (Nap) as well as fricatives (Nat), the ratio of anterior plosives and fricatives with other child articulations (%) in the maternal speech material of 2 normal children (Ch) from week 10-77.

<table>
<thead>
<tr>
<th></th>
<th>Ch</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>nap</td>
<td>136</td>
<td>591</td>
<td>1054</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nat</td>
<td>14</td>
<td>38</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nap</td>
<td>2</td>
<td>193</td>
<td>435</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naf</td>
<td>3</td>
<td>14</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% ant.</td>
<td>2</td>
<td>35</td>
<td>43</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Both mothers differed in amount in percentages in which they reacted upon the child articulations. They showed however the same tendency in their reactions upon articulation types: they reacted only upon child speech material with those articulation types which are most standard in the Dutch phoneme system. It was a remarkable fact that a high percentage of anterior plosives and fricatives were reinforced and therewith strengthened by the mothers. It looked as if they selected very carefully from all articulatory springs they heard out of the mouth of their child, those articulations which are most important for later word usage. They facilitated therewith the phonetic and phonological learning process.

In that sense the CP child with a less amount of articulations and less varied articulatory ability is not just at risk for speech and language problems due to its oral physical inability to produce sufficient anterior plosives and fricatives, but due to maternal speech interaction as well.

Remarkable differences between the 3 CP and 2 normal children were also found in onset of the vocabulary spurt, after the point of time of reached word border. The 3 CP children can be considered as delayed. (see Table 5).

Table 5. Overview of the point in time in weeks (w), on which the word border is reached in 3 CP and 2 normal born children. One child (2) had not reached this border yet at the age of 2:0 years.

<table>
<thead>
<tr>
<th>Ch.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>week</td>
<td>?4</td>
<td>80</td>
<td>53</td>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>

4.0 Conclusion

In comparison with normal born children, the cleft lip and palate children can be considered to be at risk for speech disturbances. The laryngeal expressions were more dominantly present than single articulation movements in the first two years of life. The anterior articulations were less present than the posterior ones in de CP group. The mothers of the two normal born children gave consistently verbal feedback concerning those articulation types the child uttered which belonged to the Dutch phonological system. They had high percentages of articulatory reinforcement of anterior plosives and fricatives as well. Such feedback had in our opinion an impact upon the point of time in weeks in which the word border was reached. The three CP children were far more delayed than the two normal born ones. This is of clinical importance, implying that speech rehabilitation should start already in the first year of life of cleft palate babies.

5.0. Literature

ACOUSTIC EVIDENCE THAT POSTLINGUALLY ACQUIRED DEAFNESS AFFECTS SPEECH PRODUCTION

R. Cowie, E. Douglas-Cowie and J. Rahilly

Queen's University, Belfast, Northern Ireland

ABSTRACT

31 controls and 23 speakers with severe to total acquired hearing losses were recorded reading a set passage and describing a day in their life. Samples were digitised using 1/3 octave filters. Their output was passed to programs which characterised the signal statistically. Control and deafened speakers showed a range of significant differences. Deafened speakers' average spectra showed an overall upward shift, plus over-concentration of energy and more particularly changes in 1-2 kHz. Their F0 showed an upward shift and increased spread, plus sex-dependent changes in tune shape, and their reading showed fewer stretches where F0 selected more than once. Their speech amplitude showed higher variance than controls'. Rises in amplitude were too protracted, and falls too large.

1. INTRODUCTION

This paper is concerned with the speech of deafened people, i.e. people with postlingually acquired hearing losses.

It is controversial whether acquired deafness leads to speech deterioration. Goehl and Kaufman [2] argued with some justification that studies which claimed to show speech deterioration were inconclusive for various reasons, including subjectivity of measures, small sample size, and lack of adequate controls. We report research which meets those points. It uses totally objective measures, and it compares speech from a substantial sample of deafened people with speech from a similar number of controls.

The approach was prompted by looking at spectrograms of deafened speech. It is often visually obvious that these are abnormal, but hard to pinpoint the problem in terms of local phonetic features.

This led us to develop methods which focus on gross statistical attributes of the speech signal. That approach allows us to demonstrate undeniable differences between deafened speakers and controls.

Other aspects of our work follow up and provide more detailed, linguistically oriented descriptions.

2. METHOD

The sample consisted of 54 subjects, 23 controls and 31 deafened. The deafened subjects almost all had losses over 80dB in the worse ear, so that the picture was not confused by the less severely affected speech of speakers with milder losses. Subjects were tape-recorded reading a short passage, and describing a day in their lives. This gave a range of styles from formal to a more spontaneous style.

Analysis used an ARIEL spectrum analyser housed in an IBM PC. It contains 31 filters with centre frequencies running from 20Hz to 16kHz in 1/3 octave steps, and a 32nd filter for the amplitude of the signal. A signal capture program sampled the output of these filters at 40ms intervals, and stored the results in files. Gain control was adjusted so as to use the full output range of the filters. Amplitude measures are relative to the peak amplitude in a passage (which was set to 100). Hence the analysis cannot address problems with absolute volume. But though these certainly do occur, they were not salient in our speech sample.

The analysis program takes files from the first as its input. The analysis can be thought of as involving three phases. The first extends the description of the signal. The second obtains graphs which summarise some aspect of a signal. The third extracts a range of statistical parameters which are associated with each graph.

The first phase provided four descriptions of the signal. These were the basic spectrum obtained by the filter bank, the trace of amplitude provided by the 32nd filter, and a trace of fundamental frequency. The filters are not an ideal basis for extracting fundamental frequency, but we developed a reasonably robust algorithm.

The output was always checked, and we rejected passages where we were not confident of its output. The fourth description we call a sharpened spectrum. It measures the PSI of each point in the spectrum relative to the points immediately above and below it. The value at each point is the value of the corresponding point in the basic spectrum minus a proportion of the values just above and below it.

Most of the graphs generated in the second phase are histograms. Amplitude and F0 contours were also used to generate spectrograms, mainly by plotting each point against its predecessor. This kind of treatment has interesting properties, but it led to few significant results here and so it will not be reported.

In the largest block of histograms each channel is associated with one of the frequency channels in the spectrum analyser. The simplest of these show the average level at each frequency in the basic spectrum and the sharpened spectrum, and the peak level at each frequency. More complex descriptions deal with change in the spectrum.

A set of histograms deals with sample-to-sample change. For each channel we obtain a measure which is the average (root mean square) of the differences between each value in a channel and its predecessor. This is done for both the basic and the sharpened spectrum, giving two more derived histograms.

October 1983

The histograms were derived from a measure which we call peak-to-peak change. Roughly speaking, it deals with change between successive syllable centres, whereas the sample-to-sample measure is dominated by change within syllables. The peak-to-peak measure only uses samples where overall amplitude is at a maximum. At each point the peak amplitude of each channel is compared with the value associated with the same channel at the last maximum. The differences between them are used to construct a family of histograms analogous to the histograms for sample-to-sample change.

From these descriptions another set follow. They involve the ratios of different measures in corresponding channels. For instance it is sensible to consider change/average energy; high rates of change in a channel with low average energy mean something different from similar rates in a channel where the signal is generally strong.

Histograms of a different kind were used to summarise the amplitude and F0 traces. Both were again considered on two levels, one based on point by point description and the other based on the identification of higher order structure in the trace.

For amplitude, the point by point treatment generated two histograms. In one, each channel sums the observations at a particular amplitude. In the other, it showed the number of observations which differed from their predecessor by a particular amount (being signed, not absolute differences).

Higher order structure was found by picking maxima and minima in the contour, and looking at the properties of segments which ran from a maximum to the next minimum or vice versa. Histograms were formed specifying the distributions of amplitudes at each minimum and at the corresponding maximum, the distribution of rises in amplitude between points of inflection and the distribution of falls in amplitude between all points of inflection; the distribution of the durations of rises in amplitude between points of inflection; and the distribution of the durations of falls in amplitude between all points of inflection.

For F0, the point by point treatment generated one histogram, showing the number of observations at a particular amplitude. Higher order structure involved two types of limit. The contour was divided into continuous stretches, bounded by intervals where F0 was absent. Maxima and minima were then marked on each stretch. Stretches were then assigned to one of six types: rises, falls, rises/falls, levels, falls, and compound stretches. The last type contains stretches with more than one inflection. One histogram showed the distribution of these types. A second histogram dealt with the distributions of pitch changes in segments (i.e. the interval from the highest point in each segment to the lowest). In the third phase statistical parameters were derived from each histogram. To summarise the central tendency and spread of each histogram we

350

351
Inferential statistics were applied to the measures provided by the third phase to establish where deafened and control speakers differed systematically. Unless otherwise stated all effects reported here emerged as significant effects or interactions from analyses of variance with two between variables, sex and hearing level (control or deafened); and one within variable, passage.

3.1 Spectral Abnormalities. Overall, the mean of the spectrum is shifted upwards by about 1/5 octave in the deafened speakers. The deafened also show much greater overall change in the spectrum. This is true on any measure of change. More specifically, the deafened show an abnormal concentration of change in the lower frequencies. This is shown by the significantly lower variances associated with most of the distributions of change across the spectrum.

The measures which use formant related frequency bands provide more detail.

The F2 band is anomalous on almost any measure. Among the deafened speakers the average energy there is too high, change there is too great on any criterion, and energy is more peaked at any given instant. The effect is particularly marked among females in the reading passage.

In the F1 region, the problem is more restricted to males who showed excessive rates of change. High change in this region is also consistently associated with the reading passage and with males.

There is a related problem in the F0 region, but once again it is more restricted. With one measure of change, the peak-to-peak measure, the deafened show significantly raised change relative to the absolute energy in the region. The measure is also affected by style. In the controls, change of a quite advanced energy in free speech than it is in reading. That effect is much less marked in the deaf.

At the other end of the spectrum, the fricative region shows no effect of hearing on any simple measure we used. However the deafened show a high ratio of average to peak energy in the region—that is to say the energy in that region is spread too evenly across time. That is the opposite of the kind of effect that occurred in the F0 band, at the other end of the spectrum.

3.2 The F0 contour. This topic is complicated by problems in extraction. Initially we believed that F0 was showing no large scale abnormalities, but a different picture has emerged from reanalysis using measures which are insensitive to the shortcomings of our F0 extraction. The median was taken as the most robust index of each subject's central pitch. The table below summarises average values of the medians. Both sex and hearing have significant effects.

<table>
<thead>
<tr>
<th>Table 1: Averages of subjects' median Pitch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>females</td>
</tr>
<tr>
<td>males</td>
</tr>
</tbody>
</table>

As a robust measure of pitch range we took the distance between the lowest observation and the point below which 75% of the observations lay. There is a relatively consistent pattern of increased range among the deafened, and this is mirrored in an analysis of variance which shows a marginal effect of hearing (0.1> p > 0.05).

The other abnormalities in F0 involved high order structure. The controls show a marked increase in compound features in the reading passages—that is, there are more strong peaks, which are not necessarily more peaks, but peaks which are not necessarily more peaks, but peaks which are not broken through more than one inflection. This pattern is greatly reduced in the deafened. The natural inference is that they fail to make a style shift towards more elaborate phrasing in reading.

A separate effect emerges from grouping simpler features into those which end with a fall and those which end with a rise. (Levels are ignored). A significant interaction is found between hearing, sex, and feature type. Deafened males use features which end in a rise more than any other group do, and features which end in a fall much more. It is tempting to link this to the concept of the universal of intonation. However deafened females show too many of both categories.

3.3 Amplitude. The average variance of amplitude was too high in the deaf, particularly in the reading passage. Table 2 shows how variance differs between the two groups.

<table>
<thead>
<tr>
<th>Table 2: variance of amplitude as a function of hearing and passage</th>
</tr>
</thead>
<tbody>
<tr>
<td>controls</td>
</tr>
<tr>
<td>females</td>
</tr>
<tr>
<td>males</td>
</tr>
</tbody>
</table>

High variance means that the deafened spent too little of their time at amplitudes which were near their average. Statistics concerned with maxima and minima augment the picture of how this happens.

One way of spending too much time far from the average is to oscillate between extremes. If the deafened did that, then mean amplitude at maxima should be too great and the mean amplitude at minima too low. In fact there was no significant effect of hearing on mean amplitude at either maxima or minima. Conversely both variance of amplitude at maxima and variance of amplitude at minima were too great in the deaf. Again, we would expect the opposite if the deafened were simply oscillating between loud and silent.

More detail comes from the properties of the change, the distance between maxima and minima. Overall, the variance of change per segment was too high among the deafened. However that measure combines rises and falls, and they behave quite differently. There were no significant abnormalities in the behaviour of amplitude change per rise, but both the mean duration of rises and the variance of rise duration were too high among the deaf. This is to say that rises tended to be big enough, but drawn out too long. Conversely, both the mean amplitude change per fall and its variance tended to be too high among the deafened, whereas the duration of falls showed similar means and variances for both groups. This suggests that the deafened tended to make drops in amplitude which were too big, though they lasted for the right time.

Combining these observations, only one obvious explanation for the general high variance of amplitude remains. It is that speaking style is used to accentuate both high and low extremes. This is not non-trivial, as it means that deafened speech is more appropriately characterised with terms from stylistic rather than clinical literature. That is, we may not be able to use terms like "deafness" to characterise this phenomenon.

3.4 Style shift. We have mentioned some effects which relate to style shift already. Choosing the right register is an important part of speech, and a speaker who cannot do that is faced with a non-trivial problem. There are consistent indications that deafened people have that kind of difficulty, but we will only mention a few. Among the controls, the variance of amplitude was lower in the reading passage. The deaf reversed that trend, showing slightly more variance in the reading. The controls showed a lower mean change per rise in the reading passage: that effect was minimal in the deafened. We also found style effects in non-common measuring the relationship between change in one segment and change in the next. Among the controls, these correlations were stronger in the reading passage, than in free speech — i.e. volume became less like a sequence of rises followed by similar sized falls. In the deafened, we found the opposite pattern.

4. DISCUSSION

There is a problem in the technique of using a battery of statistical descriptors to characterise speech as a distribution of energy, and we do not yet have a general domain. In this domain, it makes clear the existence of quite gross abnormalities in deafened speech. It also establishes that deafened speech shows strong common trends: it does not just shift unpredictably and idiosyncratically. This is not universally expected.

The trends which we have reported provide a focus for closer study. Since the concentration of energy around 1-2kHz emerges as a strong trend, looking at possible explanations should be a high priority, as should looking at explanations of the high variance of amplitude. The existence of problems with style shift has clear methodological implications, and since it presumably involves central control, raises theoretical issues. We are following through such questions in more detailed studies.

5. REFERENCES

MEAN-TERM PERTURBATIONS OF THE PSEUDO-PERIOD OF THE GLOTTAL WAVEFORM

R. De Guichen and J. Schoentgen

Institute of Phonetics, Free University of Brussels
Av. F.D. Roosevelt, 50, B-1000 Brussels, Belgium
* Research Associate, Belgian Fund for Scientific Research

ABSTRACT

Jitter is defined as the fluctuations from one glottis cycle to the next of the duration of the fundamental period in voiced speech sounds. We propose to study jitter from a time series point of view. Results obtained in a framework of an experiment on sustained vowels show that in a majority of cases adjacent period durations are strongly correlated and that the relationships between neighbouring periods are not the outcome of systematic long-term melodic variations.

1. INTRODUCTION

In this article we study jitter from a time series point of view. By jitter one understands the fluctuations from one cycle to the next in the duration of the fundamental period in voiced speech sounds.

Jitter has been studied for some thirty years (e.g. [4], [3], [8]). Conventionally, the amount of jitter during sustained phonation, for instance, is estimated by measuring successive glottal cycle durations and by computing a suitable dispersion measure over an appropriate interval of, typically, fifty periods. Frequently, the differences between individual periods and a long average are taken into account instead of the differences between nearest neighbours. The purpose of the running average is to remove the effects of any long-term trend on the durations of individual cycles. Trends are believed to be the consequence of melodic variations, i.e. a prosodic phenomenon under voicing control.

There are two hidden assumptions in the approach described. Both are unwarranted:

1) The first assumption is that, once any trend has been removed, the differences between adjacent periods are statistically independent. Indeed, it is under this hypothesis alone that a dispersion measure is a sufficient descriptor of jitter. In other words, neglecting the time series aspect of a sequence of cycle durations is only without consequence when periods are interchangeable.

2) The second assumption is related to the first. It consists of supposing that any systematic relationship between neighbouring periods can be effectively removed by smoothing and that, moreover, what is so removed is not words taking into account since it is simply the consequence of melodic variations.

Both hypotheses can be dropped when the sequence of glottal cycle durations is examined from a statistical time series point of view. Preliminary results we obtained thus show that the assumptions laid out above are indeed unjustified. Differences between neighbouring periods are not statistically independent since random short tern fluctuations appear to be superimposed on stochastic mean-term perturbations. This does not seem to be the consequence of a smoothly evolving melodic curve. These observations hold for a great majority of our speakers. Most of them show a positive correlation between the durations of adjacent periods. This means that longer than average durations tend to be followed by longer cycles and that the vice versa. Shorter than average durations tend to be followed by shorter cycles.

2. SUBJECTS AND METHODS

Twenty five adult speakers served as subjects for this preliminary study (eight healthy males, five healthy females, five dysphonic males and seven dysphonic females). They were told to sustain three vowels ([a], [u], and [i]), at a comfortable pitch and loudness level, as long and as steadily as they could.

The signals were recorded in a sound-proofed room. The microphone was placed approximately 5 cm from the lips. The laryngograph (or EGG for electroglosgraph) signal, which varies proportionally to laryngeal conductance, was recorded simultaneously. The signals were digitized by a two channel SONY PCM audio processor and recorded on video tape. A central one-second portion of the signal of each vowel was digitized at a 20 kHz sampling frequency with 12 bit resolution and stored for further processing in two files (EGG-and acoustic signal) on the hard disk of a Masscomp 5050 computer. An algorithm was designed to measure the duration of individual glottis cycles made use of oversampling to obtain high resolution in time. The measurements were made in two steps:

- Firstly, a gross detection of the important events in the original signals was carried out, i.e. (i) the peaks in the first derivative of the EGG signal, which were assumed to mark the instant of glottal closure, and (ii) the zero crossings in the filtered acoustic signal.
- Secondly, a portion of the signal containing the main events was oversampled eight times and low-pass filtered; the period markers were then detected with improved accuracy, leading to a numerical resolution in time of 0.625 μs. A statistical test was used to check oversampling reliability [2].

The algorithm was applied simultaneously to both the EGG and the acoustic signals. Tests carried out so far have shown that the algorithm performs satisfactorily: the comparison of the period values measured shows that both signals agree within the fine detail of the period-to-period fluctuations [7].

The algorithm also provides possibilities for graphical visualization (series of the period values, trend, differences between instantaneous period values and running averages, statistical distributions, etc...), and a battery of statistical tests. So far we have implemented five different tests (four out of five verify the statistical independence of consecutive period fluctuations, i.e. our null hypothesis):

1) The comparison to a gaussian distribution of the distribution of the microfluctuations.
2) The run test for randomness.
3) The comparison of the statistical distributions of adjacent local deviations.
4) The Pearson's moment product correlation coefficient of adjacent period durations.
5) The rank correlation coefficient of adjacent period durations.

3. RESULTS AND DISCUSSION

We have summarized in table 1 the results of serial correlation tests carried out on period sequences obtained from male and female speakers. They show that in a majority of vowel signals give rise to a positive correlation between adjacent period durations. Typical period sequences are shown in Figure 1. Figure 1a displays the period time series of male and Figure 1b the period time series of five female speakers. The first sequence in figure 1b presents a case of negative correlation between adjacent periods; all the other sequences present positive correlations.

The mechanisms underlying the production of jitter are not yet fully understood. Neuronal- and cardiac mechanisms, which have been shown to contribute to jitter [1], [5], would lead us to expect fluctuations of the fundamental period straddling several cycles. Indeed, in an enumeration of possible concurrent mechanisms, Pinto and Titze [6] distinguish between short-term and long-term contributors. Among the former they include the irregular distribution of mucus on the vocal folds, asymmetries in vocal fold geometry, turbulence, and the coupling between the glottis and the vocal tract. They count the neurological factors among the long-term aspects. What this list suggests is the existence of two time scales on the level of which independent factors are active. This point of view is
not contradicted by our preliminary findings.

On the other hand it cannot be excluded that statistical models can be shown to exist which describe the cycle duration time series purely in terms of a deterministic component driven by a purely random signal. The need to distinguish between short-term and long-term perturbations could thus be obviated.

Table 1
Results of the Pearson’s moment product and the rank correlation test for healthy and dysphonic speakers. Displayed are the number of signals showing positive correlation, no correlation or negative correlation between adjacent period durations.

<table>
<thead>
<tr>
<th></th>
<th>Speech signal</th>
<th></th>
<th>EGG signal</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson</td>
<td>Rank</td>
<td>Pearson</td>
<td>Rank</td>
</tr>
<tr>
<td></td>
<td>+ 0 -</td>
<td>+ 0 -</td>
<td>+ 0 -</td>
<td>+ 0 -</td>
</tr>
<tr>
<td>Healthy sp. (13)</td>
<td>10 1 2</td>
<td>12 0 1</td>
<td>10 1 2</td>
<td>11 1 1</td>
</tr>
<tr>
<td>Dysphonic sp. (12)</td>
<td>9 2 1</td>
<td>11 0 0</td>
<td>9 0 3</td>
<td>11 1 0</td>
</tr>
<tr>
<td>TOTAL (25)</td>
<td>19 3 3</td>
<td>23 0 2</td>
<td>19 1 5</td>
<td>22 2 1</td>
</tr>
<tr>
<td>[1]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Healthy sp. (13)</td>
<td>5 3 5</td>
<td>13 0 0</td>
<td>5 3 5</td>
<td>13 0 0</td>
</tr>
<tr>
<td>Dysphonic sp. (12)</td>
<td>9 1 2</td>
<td>11 1 0</td>
<td>10 1 1</td>
<td>12 0 2</td>
</tr>
<tr>
<td>TOTAL (25)</td>
<td>14 4 7</td>
<td>24 1 0</td>
<td>15 4 6</td>
<td>25 0 0</td>
</tr>
<tr>
<td>Healthy sp. (13)</td>
<td>11 2 0</td>
<td>13 0 0</td>
<td>11 2 0</td>
<td>13 0 0</td>
</tr>
<tr>
<td>Dysphonic sp. (12)</td>
<td>8 1 3</td>
<td>10 1 1</td>
<td>10 0 2</td>
<td>12 0 0</td>
</tr>
<tr>
<td>TOTAL (25)</td>
<td>19 3 3</td>
<td>23 1 1</td>
<td>21 2 2</td>
<td>25 0 0</td>
</tr>
</tbody>
</table>

4. REFERENCES

Figure 1
Period time series measured on a one second analysis interval for five male (fig. 1a) and five female (fig. 1b) speakers. The vertical axis is labelled in milliseconds. The average of each sequence has been offset by a constant amount, in order to avoid overlap. The horizontal axis gives the period number. The average fundamental frequencies are respectively equal to 89, 117, 123, 161 and 173 Hz in fig. 1a and to 197, 197, 200, 236 and 279 Hz in fig. 1b (from above to below). The first sequence in figure 1b presents a case of a negative correlation between neighboring periods; all the other sequences present positive correlations.
THE INTERACTION OF SPEECH PERCEPTION AND READING ABILITY

M. Gósy

Research Institute for Linguistics, Budapest, Hungary.

ABSTRACT

The preliminary hypothesis of this paper is that there should exist a close interaction between speech perception and reading ability which might result in the predictability of future reading acquisition. A tracking experiment has been carried out to support this assumption. Two classes of first-graders were examined by the GMP test-package. Data concerning their speech perception level were compared to their reading performance. Both the interaction between speech perception and reading, and the predictability of reading acquisition were confirmed.

1. INTRODUCTION

During the past few years an increasing number of children have been judged "dyslexics" because of their reading and writing difficulties. In Hungary as well as in many other countries all over the world. Experts of the problem of reading and dyslexia claim that any component of the language faculty - i.e. any of the several autonomous subsystems: phonology, syntax, or semantics - and the processing system, as well as the working memories might be the source(s) of reading difficulties [5].

As a conclusion, it has been suggested that all deficits clearly tend to co-occur (though not necessarily all), however, poor performance in terms of speech perception and understanding can always be found with poor readers. Phonetic speech perception deficits were found with American dyslexic children who had problems in the identification of places of articulation of stops and the quality of vowels. The authors' conclusion is that the deficiency is, in fact, not auditory, but a perceptual problem suggesting genetic transmission [4].

Cerebral dominance seems also to be a factor contributing to correct linguistic operations. It is likely that mixed handers might have deviations also in their language processing with regard to that of clearly right or left handers. The difference between right or left vs. the mixed handers is that the latter's two hemispheres are equally involved in linguistic behaviour. On the basis of the assumed close interaction between the speech perception/understanding process and reading ability, our hypothesis is that reading performance is predictable.

2. PROCEDURE

At the Phonetics Laboratory in Budapest a special test-package (GMP) has been set up in order to detect children's ability for actual reading and for future reading acquisition [3].

In compiling the test-package, efforts have also been made to obtain information on the operations of each hypothetical level of the speech perception process quasi-separately, i.e. to detect which (if any) of the decisions the understanding mechanism has to perform are mistaken or incorrect.

The GMP test-package consists of 14 subtests: their naturally announced and artificially generated synthesized speech material varies from isolated words through sentences up to a longer text. These speech materials have been manipulated by various methods (such as masking by white noise, speeding up, and frequency filtration). Some of the listening tests have been administered to the subjects through headphones, others through a loudspeaker in a silent room. The sub-tests measure both peripheral and central hearing, acoustic, phonetic, phonological levels of speech perception, visual and verbal short-term memory performance, lip-reading ability, handedness, directions, repetition ability of speech rhythm, word-completion skill, and text-comprehension.

300 normal hearing children (ages between 5 and 8) have been examined with the test-package in order to define age-specific values for normal performance. Figure 1 shows the developmental results of the GMP subtests. The examination with the GMP test-package takes about 30 minutes, both the (kindergarten/school) teachers and the speech therapists can use it easily. 150 children suffering from reading difficulties were also examined by means of the GMP. On the basis of the results the reason(s) of their reading difficulties could be detected on the one hand, and a corrective therapy could be proposed on the other. The re-examinations confirmed that the diagnosis was correct.

3. RESULTS

A tracking experiment has been carried out to support the predictability of somebody being a poor reader. 37 first-graders (21 girls and 16 boys) participated in this experiment who learned in the same school but in two separate classes. (Their sociological background was very similar.) The children have been examined by the GMP test-package at the beginning of their first school-year and they have been re-examined after 4 months. During this time they were taught by the same teaching method, books etc. (Efforts have been made to choose similar personalities as their teachers.) By the end of this 4-month period the children had to know all Hungarian letters (both in reading and writing) and had to be able to read simple sentences correctly. At the end of this period, the same Reading Assessment Test (RAT) has been carried out with the children in order to check their reading level. There was no significant difference in the GMP results of the two classes at the first examination (Table 1) while there were highly significant differences among the children (p<0.01).

15 children (7 from Class A and 8 from Class B) have been found pronouncing metatheses while repeating the meaningless sound sequences, and 18 children (8 from Class A and 10 from Class B) suffering from direction disturbances. Left-ear-advantage was found with two children. There were 11 children (5 from Class A and
The children’s data show various co-occurrences of problems as shown by the GMP-subtests, such as a mixed-bander pronouncing metatheses, having problems in identifying the speeded-up sentences, or a righthander with no articulation problem, normal speech perception performance but poor verbal short-term memory and poor text-comprehension. Which of these co-occurrences can significantly predict the poor reading performance? Our basic hypothesis is that those children should be judged as possible poor readers who (i) show a poorer performance in (almost) every subtest of the GMP than it is required for their age level, (ii) have poorer performance in more than two subtests, and (iii) have an extremely poor performance in one of the subtests, particularly in the identification of fast sentences. But the latter 5% of their GMP results which were significantly poorer than that of others (p<0.001), 12 children (5 from Class A and 7 from Class B) were predicted to have difficulties in reading acquisition.

The children’s performance with the GMP test-package shows significant differences between the two classes at the second examination, similarly to reading performance (p<0.05). The results are significantly better in Class A where the special corrective course was performed. Data obtained in subtests for understanding of reading show a larger difference between the two classes (p<0.01). Table 3 contains our predictions concerning children’s expected reading acquisition and their confirmation in terms of RAT results.

The distribution of children in terms of RAT performance shows greater diversity in Class B where no corrective course was conducted than in Class A (Table 4).

### Table 3

<table>
<thead>
<tr>
<th>Predictions and supporting data on reading ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Average GMP Perform. in readers results (points) RAT (points)</td>
</tr>
<tr>
<td>'good' 88.6</td>
</tr>
</tbody>
</table>
| 'poor' 65.1 | 90-96*
| 'poor' 66.3 | 65-85** |

* (after corrective course)  
** (without corrective course)

### Table 4

<table>
<thead>
<tr>
<th>Distribution of children according to their results in reading test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class A</td>
</tr>
<tr>
<td>Points</td>
</tr>
<tr>
<td>100</td>
</tr>
<tr>
<td>95-99</td>
</tr>
<tr>
<td>90-94</td>
</tr>
<tr>
<td>85-89</td>
</tr>
<tr>
<td>80-84</td>
</tr>
<tr>
<td>75-79</td>
</tr>
<tr>
<td>70-74</td>
</tr>
<tr>
<td>65-69</td>
</tr>
</tbody>
</table>

Two important conclusions can be briefly drawn.

1. Speech perception and comprehension performance shows a very close interaction with reading ability. It is not only the operations at the hypothetical levels of the speech understanding mechanism that should be taken into consideration, but also the concomitant abilities and capabilities of children. There is a high correlation between their performance in these two tasks and their reading performance.

2. Reading ability can be assessed before the children begin to learn reading and writing, i.e. reading performance is predictable. The majority of children’s problems in relation to language and particularly speech perception should be compensated for in preschool age. This offers a good prognosis for successful reading acquisition.

### 6. REFERENCES

ANALYSIS OF GLOTTAL WAVEFORM IN DIFFERENT PHONATION TYPES USING THE NEW IAIF-METHOD

P. Alku1, E. Vilkman2, U.K. Laine1
1: Helsinki University of Technology, FINLAND
2: University of Oulu, FINLAND

ABSTRACT
A new glottal wave analysis method, IAIF (Iterative Adaptive Inverse Filtering), is presented. In this algorithm the effect of glottal pulses to the speech spectrum is first estimated with an iterative procedure. The model for the vocal tract is then computed using linear prediction. Finally, the effects of the vocal tract and lip radiation are cancelled by inverse filtering. The IAIF-method was tested using synthetic and natural vowels of three different phonation types. The new algorithm was able to yield a fairly accurate estimate for the glottal excitation excluding the case of very pressed phonation that was partly distorted by a formant ripple.

1. INTRODUCTION
Many different methods have been developed during the last forty years in order to estimate the source of voiced speech, the glottal pulseform. One of the methods that are commonly used in the analysis of the glottal excitation is inverse filtering. Although good results have been obtained with this method it has some drawbacks. For instance, the transfer function of the inverse filter is often adjusted manually. Hence, the final result is very much dependent on the subjective criteria applied by the researcher [e.g. 3]. Another drawback that is characteristic especially to the closed phase covariance method is that the analysis works properly only for phonation types with a sufficiently long glottal closed phase.

In this paper a new glottal wave analysis method, IAIF (Iterative Adaptive Inverse Filtering), is presented. The method represents further development of the IAIF-method, which has been presented earlier [1]. Therefore a brief description of the algorithm will be given in section 2. The performance of the IAIF-method in the estimation of the glottal excitation is discussed in section 3 using both synthetic and natural utterances.

2. METHOD
The IAIF-method is based on a speech production model that consists of three separated processes: the glottal excitation, the vocal tract and the lip radiation effect. The model is assumed to be linear and the interaction between the three parts is considered to be negligible. The vocal tract is modeled with an all-pole filter. The last process of the model, the lip radiation effect, is modeled with a differentiator.

The block diagram of the IAIF-method is shown in Fig. 1. The speech signal to be analysed is denoted s(n) and the result, the estimate for the glottal excitation, is denoted g(n). The first iteration consists of the blocks numbered from 2 to 6 and the second iteration of the blocks numbered from 7 to 11. The purpose of each of the blocks is described as follows.

Block no. 2: The effect of the glottis to the speech spectrum is preliminarily estimated by first order LPC-analysis.

Block no. 3: The estimated glottal contribution is eliminated by filtering $s_{hp}(n)$ through $H_{p}(z)$.

Block no. 4: The first estimate for the vocal tract is computed by applying LPC-analysis to the output of the previous block.

Block no. 5: The effect of the vocal tract is eliminated from signal $s_{hp}(n)$ by inverse filtering.

Block no. 6: The first estimate for the glottal excitation $g_1(n)$, is obtained by cancelling the lip radiation effect by integrating.

Block no. 7: The second iteration starts by computing a new estimate for the effect of the glottis to the speech spectrum. This time second order LPC-analysis is used. The signal from which the glottal contribution is eliminated is $g_2(n)$.

Block no. 8: The effect of the estimated glottal contribution is eliminated.

Block no. 9: The final model for the vocal tract is obtained by applying LPC-analysis of order $r$ to the output of the previous block.

Block no. 10: The effect of the vocal tract is eliminated from speech by filtering $s_{hp}(n)$ through $H_{v}(z)$.

Block no. 11: The result, $g(n)$, is obtained by cancelling the lip radiation effect by integrating the output of block no. 10.

The results discussed in this paper are based on the implementation of the IAIF-algorithm on a Symbolics Lisp-machine.

3. RESULTS
3.1. Synthetic vowels
In order to verify the performance of the IAIF-method the new algorithm was first tested with synthetic speech. Synthetic vowels were created using a procedure described in [4]. The vocal tract was modeled with an eighth order all-pole filter and the lip radiation effect with a differentiator. The shape of the vocal tract transfer function corresponded to the vowel /a/. The signal bandwidth was 4 kHz. As the synthetic source signal we used a glottal pulse model described in [2]. Three different phonation types, breathy, normal and pressed, were simulated by changing the shape of the synthetic excitation waveform. Two different values for the pitch period, corresponding to male and female speakers, were used in the synthesis procedure.

The IAIF-analysis was computed for all the signals using a block length of 256 samples (32 ms). The orders of LPC-analysis corresponding to modeling of the vocal tract (parameters $p$ and $r$ of blocks no. 4 and 9 in Fig. 1) were chosen to be equal. This value was varied from 8 to 12 by a step of two.

When synthetic male phonation was analysed the IAIF-method yielded a result.
that was very close to the original source signal. In the case of breathy and normal phonation similarity between the original source and the waveform given by the IAIF-method was almost exact without dependence on parameter p. A typical result is shown in Fig. 2. In the case of pressed phonation the waveform obtained by the IAIF-method was partly distorted by a ripple component when the value of p was equal to 8 i.e. to the order of the all-pole vocal tract. However, by increasing the order of p to be equal to 12 the ripple component disappeared.

![Fig. 2. Analysis of a synthetic male vowel of breathy phonation (a): Original synthetic glottal source (b): Glottal wave estimate given by the IAIF-method (p=8)](image)

When synthetic female utterances were analysed the results were not so good as for male voices. In the case of breathy phonation the IAIF-method gave a waveform that was similar to the synthetic source signal. However, for normal and in particular for pressed phonation types the result given by the new algorithm was partly distorted by a ripple component. This results from the spectrum of the glottal excitation which in the case of pressed phonation comprises more high frequency components than in the case of breathy or normal phonation. In the case of female voice the source spectrum is also characterized by a sparse harmonic structure. Hence, LPC-analysis (block no. 9 system 1) gives a vocal tract filter, where the formants, especially F1, are moved from their original positions because of the harmonics of the source spectrum. Thus, a small formant ripple will be present in the glottal wave estimate after inverse filtering and integration.

![Fig. 3. Glottal wave estimate given by the IAIF-method (natural male voice, p=12) (a): Breathy phonation (b): Normal phonation (c): Pressed phonation](image)

3.2. Natural vowels

The IAIF-method was used in the glottal wave analysis of sustained phonation by studying utterances that were produced by one female and one male speaker. Both of the subjects were of healthy voice. The speakers were asked to produce the vowel /a/ using breathy, normal and pressed phonation. The recording was done in an anechoic chamber using a condenser microphone (Driel&Kjær 4134). The speech material was A/D-converted with Sony PCM-F1 and stored on a video cassette using Sony SL-F1E. The bandwidth of the signals was downsampling to 4 kHz.

In the case of male voice the results were of reliable shapes for breathy and normal phonation types. The glottal waveform corresponding to pressed phonation was partly distorted by a formant ripple. Fig. 3 shows the obtained glottal pulseforms for all the three phonation types.

![Fig. 4. Glottal wave estimates given by the IAIF-method (natural female voice, p=12) (a): Breathy phonation (b): Normal phonation (c): Pressed phonation](image)

4. DISCUSSION

In this paper a new glottal wave analysis tool, the IAIF-method, was presented. The identification of the different processes of the human speech production mechanism is done in the new algorithm using a frequency domain approach. The average glottal contribution to the speech spectrum is first estimated with an iterative procedure. The vocal tract is then identified by LPC-analysis. The estimate for the glottal excitation is finally obtained by cancelling the effects of the vocal tract and lip radiation by inverse filtering.

The new IAIF-algorithm was applied in this study for the glottal wave analysis using three different phonation types. The results obtained are in line with those reported using other methods [e.g. 3].

In the case of male voice both synthetic and natural utterances gave the same result: breathy and normal phonation can be analysed accurately whereas pressed phonation is partly distorted by a formant ripple. The reason for distortion with the IAIF-method was obviously the poor estimation of the first formant, which comes from the contribution of the source spectrum. For synthetic female voices, especially in the case of pressed phonation, distortion was largest. However, in general, excluding the very pressed phonation type, the source spectrum of natural female phonation decays so fast that the first formant can be modeled properly. This explains why the analysis results obtained from the utterances of the female subject were of reliable shapes with no formant ripple.

The IAIF-method has proved to be a promising tool for glottal wave analysis. The main advantage of the new algorithm is that it is automatic. Hence, the glottal pulseform can be obtained without manual interference by the investigator. Further studies are needed to compare the IAIF-technique with traditional methods as well as to reveal whether it can be used for analysis of other speech sounds. Also the real-time implementation of the algorithm using the TMS320C30-signal processor is under development.

References:


A VOICE CONVERSION METHOD AND ITS APPLICATION TO PATHOLOGICAL VOICES

H. Kuwabara¹ and T. Takagi²

¹ The Nishi-Tokyo University, Yamanashi, Japan
² Nihon Science & Tech. Res. Labs., Tokyo, Japan

ABSTRACT

Formant and pitch frequencies are used as the acoustic parameters to be manipulated. These acoustic parameters are first extracted from a speech sound to be modified and changed according to some rules that are to make the original speech clear, and a new speech is synthesized using the modified acoustic parameters. Speech intelligibility is found to reach the maximum when the trajectories are emphasized to some extent. It is also found that our method is capable of improving the so-called "roughness" or "hoarseness" of pathological voices mainly by replacing pitch frequency of the original speech with that of a normal speaker.

1. INTRODUCTION

Using the analysis-synthesis system we have developed [1], voice quality of natural speech has been controlled by changing formant trajectories that are supposed to have a close relation to such voice qualities as intelligibility, clearness and so on. Correlation analysis between psychological and acoustic distances reveals that the formant trajectories and correlation with the voice quality of the announcer's speech sounds, followed by pitch frequency [2]. This result suggests that the quality of speech sound of non-professional speakers may possibly be improved by altering the dynamics of formant trajectories.

2. ANALYSIS-SYNTHESIS SYSTEM

Fig. 1 illustrates the block diagram of the analysis-synthesis system. Low-pass filtered input speech was digitized in 12 bits at a rate of 15 kHz. A short-time LPC analysis based on the autocorrelation method was performed to obtain LPC coefficients and the residual signal. Formant frequencies and bandwidths were estimated by solving a polynomial equation. A modification of the spectral envelope was equivalent to a manipulation of the coefficients that would result in a frequency response of the filter equal to the modified envelope. These acoustic parameters (pitch periods, LPC coefficients, formant frequencies, bandwidths, residual signals) were stored for later synthesis.

3. METHOD OF FORMANT TRAJECTORY MANIPULATION

After extracting formant trajectories using the method proposed by Kasuya [4], modification of the trajectories makes it possible to alter the formant pattern in any way. The method for altering the formant pattern is the same as that we have proposed earlier for the modification of vowels in connected speech [3].

In this study, T=150ms and ω=52 ms were experimentally decided. Given d=0, the dynamics of the original formant trajectory is emphasized. While for d<0, it becomes deemphasized.

Equation (1) is applied to each of the three formant trajectories without voiceless consonant (except for voiceless consonant). The time interval in equation (1) during which the weighted sum is calculated, is 150ms, a 300ms for backward each.

This is the result for d=7.3 which, in previous study, represents a proper value for the purpose of normalizing coarticulation.
lation effects of vowels in continuous speech. It is noticed from the figure that the new formants are emphasized in comparison to those of the raw formants.

4. METHOD OF PITCH MANIPULATION

Pitch frequency manipulation is quite simple as depicted in Fig. 2. At the pitch synchronous analysis stage, the residue signal obtained for each pitch period has exactly the same data length as the pitch period. If we give the residue signal as an input to the vocal tract model, exactly the same waveform as the original speech will be obtained. Thus, pitch frequency change can basically be given by controlling the length of the residue signal.

To raise pitch frequency, some data at the last part of the residue signal is deleted and to lower the frequency, zero signals are added to the last part of the residue signal.

5. ENHANCEMENT OF PATHOLOGICAL SPEECH

An attempt has been performed to improve the quality of a pathological speech using the analysis-synthesis system we have developed. The pathological speech used in this experiment is an utterance of a patient who has a disease in his vocal cord. Because of malfunction of the vocal cord vibration, the resultant speech wave lacks clear periodicity and its voice quality is "hoarse". The experiment has been designed to create the fundamental frequencies into the pathological speech wave in order to improve the quality as close as normal speech.

Fig. 3 represents the block diagram to improve the quality of pathological speech. It requires two kinds of input speech: a pathological speech to be improved and a normal speech utterance of the same sentence from another speaker. From the pathological speech inputted, voiced portions are first detected and the spectral envelopes are extracted by LPC analysis. Then, the normal speech is analyzed by the same method and the pitch frequencies are detected to combine with the spectral information extracted from the pathological speech. If the normal speech of the same content can not immediately be available, artificial pulse trains could be used as a voice source. In the analysis stage, after making voiced/voiceless distinction, the voiceless portions (voiceless consonants and de-vocalized vowels) are thoroughly kept in memory and the LPC analysis is performed for the voiced portions to obtain LPC coefficients that carry spectral information and the residual signals from which pitch periods can be estimated. For the pathological speech, the frame length (analysis window) is set at 20 ms and the frame shift is a half the window length.

In the feature extraction stage, the residual signals for the pathological speech are discarded after obtaining spectral information. Contrary to this, only the pitch frequency contour is needed from the normal speech.

For the normal speech, however, a process of time alignment has been undertaken before feeding to analysis in Fig. 3. This process is shown in Fig. 4. The voiced parts of the normal speech are analyzed pitch synchronously and the length for each part is compared with the corresponding part for the pathological speech in order to make the length equal to that of the pathological speech with accuracy of less than one pitch period. This has been done simply by eliminating or inserting additional pitch periods.

The normal speech, after being time-aligned, is LPC analyzed again and the pitch frequencies are extracted for every voiced portion. This pitch frequencies or the residual signals are fed into the synthesis filter as the voice source. The synthesis filter is made from the predictor coefficients obtained from the pathological speech. The resultant output speech has, therefore, the same spectral characteristics as the pathological speech and the same source characteristics as the normal speech. Fig. 5 depicts an example of speech waveform for the pathological speech synthesized by the proposed method and also synthesized speech with an artificial pulse train as the voice source to the filter.

As far as we have tested, the quality of the synthesized speech has been found to be far better than the original pathological speech, though it is not

6. CONCLUSION

Improvement of voice quality has been performed using an analysis-synthesis system capable of modifying pitch, formant frequencies and formant bandwidths. According to the results of analysis for professional announcers speech sounds, it is obvious that speech intelligibility of pathological speech is due to the dynamics of formant and pitch patterns. It has been found to be possible to improve the speech intelligibility without changing voice individuality by emphasizing the movement of time-varying pitch pattern. Another application of this analysis-synthesis system has been made to enhance a pathological speech which has little periodicity and "hoarse" in voice quality. By adding fundamental frequency component taken from a normal speaker, the voice quality of the pathological speech has been improved to a great extent.

7. REFERENCES

Consistency in /r/ Trajectories in American English
Carol Y. Espy-Wilson
Electrical, Computer and Systems Engineering, Boston University, Boston, MA, USA
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT
We discuss the results of an acoustic study of the influence of postvocalic /r/ on
neighboring segments in American English. The data suggest that a certain
amount of time is needed to articulate an /r/ and that different speakers begin to
produce /r/ at different times, depending on rate and context.

1. INTRODUCTION
A salient acoustic characteristic of American English /r/ is a low third
formant (F3) which is close in frequency to the second formant (F2). F3 for /r/
occurs in the [speakers], and exceeds or is usually above 2000 Hz. There is
substantial downward movement in F3 from a canonically articulated /r/ to an /r/
in the word "car." However, in words like "cart" and "carwash," other F3 trajectories
sometimes occur where the downward F3 movement is seen earlier [1]. In this
study, we investigate the effects of speaking rate, context and speaker
differences on the articulation of /r/.

2. CORPUS
To conduct this acoustic study, the words "cart," "car," "carwash," "card," and
"carp" were embedded in the carrier phrase "Say ______ for me" and spoken by
six speakers, four females (AF, LW, LT, and MH) and two males (MR and JR). As
a neutral case, the speakers also said the word "Nadav" (/Nodov/) in the sentence
"Nadav was here." The speakers were
recorded in a quiet room and instructed to
speak at a slow and fast rate. The utterances were low pass filtered at 4800
Hz, sampled at 10 kHz and
preemphasized. F3 tracks were obtained from DFT and LPC spectra with a 25.6
ms Hamming window.

3. ANALYSIS
In this section, we present measurements of speaking rate, a
characterization of the F3 trajectories and a measure of how early speakers
anticipate the /r/.

3.1 DURATION
Measurements of the sonorant interval showed that the average /r/ duration
across speakers (except for subject JR) was 295 ms for the words spoken at the
slow rate and 182 ms for the words spoken at the fast rate. (JR did not show
dramatic differences.) As expected, average /r/ durations were longer before voiced consonants (293 ms
- slow, 221 ms - fast) than before unvoiced consonants (170 ms - slow, 142
ms -fast).

3.2 F3 trajectories
F3 trajectories observable during the sonorant region had four basic shapes.
These shapes are shown schematically in Figure 1 with spectrograms for different
pronunciations of "cart" which illustrate the corresponding F3 trajectory. First, as
shown in part (a), F3 can start from a high
position and move to a lower position (L-like). In this case, the vowel and /r/
appear to be produced canonically, with the /r/ articulation appearing at the end of
the sonorant region. In part (b), F3 is rather flat and at a low position throughout.
In this case, the /r/ and vowel appear to be completely coarticulated. In part (c), F3 moves from a low position at
the beginning of the sonorant region to a higher position towards the end (J-like).
Thus, as in part (b), it appears as if the /r/ is coarticulated with the vowel, however
movement away from the /r/ to the

"can" is still being articulated during the following /l/.

3.3 Anticipation of /r/
To develop a criterion by which it can safely be said that the /r/ is being
produced, we used the F3 minimum (Fm) during the neutral case, the /r/ in "Nadav.
" The beginning of r-coloring in the test words was taken as the time (TR) at which F3 during the test word fell 500 Hz
below Fm. The difference of 500 Hz was
chosen since other factors which can lower F3 such as the influence of a labial consonant should not result in such a
large change. To measure when speakers started to produce an unambiguously r-
-colored sound, we subtracted B from TR, the time at which the sonorant region
began. This difference was divided by the total duration of the sonorant region to
normalize for speaking rate. Thus, the
resulting values lie between 0 and 1. If F3 is 500 Hz below Fm at the beginning of the
sonorant region, the normalized difference

Table 1. Shapes of F3 trajectories across all speakers as a function of rate and context.

<table>
<thead>
<tr>
<th>Shape of F3 Trajectories</th>
<th>Subjects</th>
<th>L-like</th>
<th>Flat</th>
<th>J-like</th>
<th>U-like</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>V,P</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>all</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MR</td>
<td>all</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>JR</td>
<td>all</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>AF</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LT</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LT</td>
<td>all</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LW</td>
<td>V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LW</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>JR</td>
<td>all</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>JR</td>
<td>all</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. A comparison of F3 trajectories occurring for subjects LN and LW.
following consonant is also evident. Finally, as in part (a), the F3 trajectory of part (d) starts from a high position and drops to a minimum. However, as in part (b), F3 rises towards the end of the sonorant region (U-like).

As we will discuss below, these data support the possibility that the U-like F3 trajectory occurs in all cases; however, there appear to be differences because the full F3 trajectory does not always occur within the sonorant region where the formants are visible. The other cases can be derived from the U-like trajectory. In the case of the L-like trajectory, the latter part of the F3 trajectory is coarticulated with the final consonant so that the upward F3 movement from the F3 minimum is not visible. For the flat trajectory, the beginning and end of the full F3 trajectory occur outside the sonorant region so that only the region around the F3 minimum is visible. Finally, for the J-like trajectory, anticipation of the /r/ occurs during the initial consonant so that the downward F3 movement occurs during the aspiration noise.

Figure 2 shows three F3 measurements for each word: the beginning of the sonorant region (B), the end of the sonorant region (E) and the F3 minimum (L) for subjects LT (left) and LW (right). The upper two trajectories in each graph are measurements of F3 during the /r/ in the fast and slow pronunciations of "Nadar" which serves as the neutral case.

The plots for subject LT illustrate that some speakers have fairly uniform behavior across rate and context. The F3 trajectory is always relatively flat. On the other hand, other speakers like subject LW show more variability.

The shape of the F3 trajectories as a function of rate and context are summarized in Table 1 for each speaker. The data show that different speakers have different tendencies for when they begin to produce the /r/. For example, the speaking at a faster rate (the exception for subject AF is "carp").

4. CONCLUSIONS
The data in this study suggests that 1) the articulation of /r/ requires a minimum time of execution and 2) the acoustic consequence of the articulation of /r/ is a downward movement in F3 into the /r/ and an upward movement in F3 away from the /r/. However, this full F3 trajectory is not always observable because the formants are generally visible only during the sonorant regions of the

Figure 3. A comparison of the /r/ bursts in the slow (dotted) and fast (solid) pronunciations of "cart" by subject LW.

Figure 4. A plot of the normalized difference TR-B for subjects LW and AF

Table 1: Normalized difference TR-B for subjects LW and AF

<table>
<thead>
<tr>
<th>Subject</th>
<th>Slow</th>
<th>Fast</th>
</tr>
</thead>
<tbody>
<tr>
<td>LW</td>
<td>0.45</td>
<td>0.35</td>
</tr>
<tr>
<td>AF</td>
<td>0.50</td>
<td>0.60</td>
</tr>
</tbody>
</table>

5. ACKNOWLEDGMENTS
This work was supported in part by NSF Grant BNS-8920470. I also want to acknowledge Shawn Williams who made many of the measurements during her bachelor's thesis and Suzanne Boyce who gave many helpful comments.

6. REFERENCES
LA VARIABILITÉ INTER-LOCUTEUR,
ETUDE SUR LES REALISATIONS ACOUSTIQUES DE /e/, /e/,
A. Bonneau

CRIN-INRIA, Nancy, France.

ABSTRACT
This paper deals with the different sources of acoustic variability and particularly with across-speaker variation. We propose a speaker’s normalization procedure based on a minimal training and implementing expert knowledge. We test our procedure on two French vowels (/e/, /e/) uttered by 13 male speakers.

INTRODUCTION
Nous présentons ici une étude des variations qui agissent simultanément sur la parole afin de déterminer leurs manifestations acoustiques et leurs influences réciproques. Nous avons choisi de nous concentrer essentiellement sur la variation inter-locuteur et sur l’influence du contexte sur celle-ci.

La variation inter-locuteur est décomposée en variation d’origine physiologique et en variation d’origine articulaire. Nous proposons une méthode d’évaluation de la variation d’origine physiologique, fondée sur un apprentissage minimal grâce à l’apport de connaissances.

Nous avons choisi de commencer notre étude par Analyse de deux voyelles antérieures du français, /e/ et /e/, pour deux principales raisons: d’une part /e/ constitue une bonne voyelle d’apprentissage et d’autre part il nous semble physiologiquement intéressant d’étudier deux voyelles dont l’opposition est neutralisée en français.

1. VARIATION INTER-LOCUTEUR
Nous serons d’évaluer les effets de la variation inter-locuteur d’origine physiologique et nous évoluerons les conséquences de la variation inter-locuteur d’origine articulaire.

1.1 Normalisation et de normalisation de la variation d’origine physiologique
Nous désirons déterminer les répercussions acoustiques des différences physiologiques entre locuteurs ou, plus précisément, déterminer les variations des fréquences formantiques en fonction de la taille du conduit vocal.

Si on multipliait toutes les dimensions du conduit vocal par un même facteur, les fréquences formantiques caractéristiques de ce conduit vocal, abstraction faite des conséquences de la radiation aux lèvres, seraient également multipliées par le même facteur, l’inverse du premier.

Le rapport moyen entre les fréquences formantiques des locutrices et celles des locuteurs, toutes voyelles et tous formants confondus, est de 1.17 ce qui correspond approximativement à une différence de taille de 3 cm entre les deux conduits vocaux des deux sexes.

Les écarts observés ne sont pas les mêmes quels que soient le formant et la voyelle considérés. Ainsi selon les calculs de Peterson et Barney [1], soit Fh et Ff représentant respectivement les moyennes masculines et féminines, Ffh est égal à 1.01 pour le premier formant de /u/ et 1.23 pour le troisième formant de cette même voyelle.

La principale explication de ce phénomène est l’existence de différences physiologiques entre les configurations vocales des hommes et celles des femmes. En effet, la longueur du pharynx (Fatt [2]) et l’ouverture relative au point de constriction maximale (Traunmüller [3]) sont relativement plus grandes chez les hommes.

Une évaluation des variations fréquentielles liées aux différences physiologiques nécessiterait la réponse à la question suivante: comment s’effectue le passage d’une configuration vocale masculine typique à une configuration vocale féminine typique? S’effectue-t-il de manière continue, en fonction de la taille du conduit vocal, que le locuteur soit féminin ou masculin? Ou de manière discontinue, avec une rupture du continuum à la rose, causée par la baisse du larynx chez les hommes (Traunmüller [3]). Cette dernière solution nous apparaît la plus plausible, mais nous analyserons néanmoins les deux eventualités. Remarquons auparavant qu’il est évident que ce problème est également le problème de la portée de la normalisation qui est posé ici: s’appliquera-t-elle à tous les locuteurs, ou doit-on normaliser séparément les représentations vocales des hommes et celles des femmes?

1.1.1 Normalisation formantique conjointe des locuteurs et des locutrices
Nous faisons l’hypothèse que la longueur relative du pharynx et l’ouverture relative au point de constriction maximale dépendent uniquement de la taille du conduit vocal. Supposons que les fréquences formantiques varient indépendamment en fonction de cette taille et que les différences physiologiques constituent l’unique source de variations formantiques. Cette hypothèse peut être simplement évaluée par la position relative des fréquences formantiques d’un locuteur par rapport aux fréquences formantiques moyennes des hommes et des femmes.

Cette position relative est identique pour tous les formants de toutes les voyelles d’un locuteur puisqu’elle indique, selon nos hypothèses, la taille de son conduit vocal.

A partir d’une seule fréquence formantique d’une seule voyelle d’un locuteur, on doit prédire les fréquences formantiques de toutes les voyelles de ce locuteur.

Mais d’autres facteurs de variations fréquentielles sont à prendre en considération, comme l’articulation spécifique à chaque locuteur, que nous ignorons donc que les conséquences acoustiques sont plus difficiles à évaluer.

Nous laissons de côté pour l’instant les variations dues à la vitesse d’élocution. Etant données nos précédentes hypothèses (continuum des configurations vocales, linéarité des variations fréquentielles), les représentations acoustiques de chaque voyelle V dans le plan ou l’espace formantique se répartissent, sous l’effet des différences physiologiques, le long d’une droite définie par les moyennes masculines et féminines de V. Appelons chacune des droites ainsi définies, il en existe une par voyelle, droite “physiologique”. La méthode la plus simple pour évaluer l’effet des différences physiologiques en dépit des conséquences acoustiques des autres sources de variation consiste à effectuer une projection pérpendiculaire d’une image acoustique d’une voyelle V sur la droite “physiologique” de V et à se reporter au crête de la position relative en le modifiant légèrement puisque deux formants au moins sont désormais nécessaires pour notre évaluation.

Cette méthode, proposée par V. Lonchamp (Bonneau [4]), suppose que l’effet des variations articulatoires est négligeable le long des droites “physiologiques”.

La méthode d’évaluation exposée ici nous permet de proposer une procédure de normalisation des locuteurs féminins et masculins fondée sur un apprentissage minimal.

Partir des fréquences formantiques d’une voyelle V prononcée par un locuteur 1 ainsi que des fréquences formantiques moyennes des hommest et des femmes de cette même voyelle V, nous déterminons un paramètre de normalisation appelons-le “dix”, qui tient compte implicitement de la taille du conduit vocal de 1. Détailons la procédure:

- nous énumérons trois répétitions de /e/ et /e/ en contexte labial, émises par un locuteur 1.

En théorie, une seule répétition de /e/ et suffit mais trois répétitions au moins sont nécessaires afin de minimiser les erreurs dues à une prononciation aléatoire. 1e semble être une bonne voyelle d’apprentissage (Bonneau [4]).

- nous projetons l’image acoustique de /e/ et /e/ sur la moyenne des trois répétitions, la droite qui relie les moyennes masculines et féminines (Fm) et féminines (Ff) de /e/ et /e/.

- nous calculons le paramètre de normalisation physiologique “dix” qui indique la position relative de notre projection par rapport à Fm et Ff.

- nous voyelles de /e/ et /e/ sont normalisées par l’application de la
formule: soit \( F_{ij} \) représentant la fréquence du ième formant de la voyelle \( j \) prononcée par le locuteur \( i \) et \( F_{ij} \) sa normalisation.

\[ F_{ij} = F_{ij} - \text{diff} \]

Notre normalisation consiste donc à déplacer chaque représentation acoustique d'une voyelle quelconque parallèlement à la droite "physiologique" qui correspond à cette voyelle.

1.1.2 Normalisation formantique séparée des locuteurs et des locutrices

Une normalisation commune des locuteurs et des locutrices n'est pas possible. Si on se reporte au cas simple évoqué plus haut - c.à.d. pour une configuration vocale donnée, toutes les dimensions du conduit vocal et les fréquences formantiques sont multipliées par le même facteur quand la taille du conduit vocal varie-, on peut évaluer simplement les variations fréquentielles des locuteurs masculins à partir des moyennes formantiques masculines, et celles des locutrices à partir des moyennes formantiques féminines. Afin de limiter les erreurs d'évaluation causées par l'articulation spécifique à chaque locuteur, nous proposons de suivre une procédure d'évaluation semblable à la précédente, c.à.d. qui comporte une projection des données d'apprentissage sur une droite "physiologique". Deux droites "physiologiques" sont ici nécessaires par \( F_i \)

![Figure 1](image.png)

**Determination du paramètre de normalisation**

- axe moyen masculin de \( e \)
- rfe moyenne masculine de \( e \): représentation acoustique de cette voyelle pour le locuteur sur lequel on effectue la normalisation, \( F_e \) sa projection sur la droite des moyennes.

- voyelle, une pour les références féminines et une pour les références masculines.

Il se peut que les différences physiologiques n'aient pas des conséquences aussi triviales et qu'il faille réévaluer celles-ci. Si nous conservons l'hypothèse simple d'une variation linéaire des fréquences formantiques en fonction de la taille du conduit vocal, nous devons proposer de meilleures droites "physiologiques". Cette tâche soulève quelques problèmes que nous n'avons pas la place d'évoquer ici.

1.2 Sources d'erreurs

Ce qui précède est une version simplifiée des conséquences probables de la variation inter-locuteur. Citons quelques phénomènes qui peuvent perturber, selon leur ampleur, la bonne estimation de nos paramètres de normalisation:

- la compensation articulatorique qui affecte notamment l'ouverture relative du conduit vocal,
- l'influence très forte de certains contextes, par exemple l'influence des dentales sur l'articulation de /l/, qui peut mécaniquement déterminer la qualité des droites "physiologiques",
- les variations fréquentielles entre locuteurs d'origine articulateurale, si elles ne sont pas négligeables dans le long de nos droites physiologiques, et qui sont d'autant plus difficiles à cerner qu'elles peuvent changer pour un même locuteur avec le contexte.

Le bien-fondé des procédures de normalisation de la variation inter-locuteur d'origine physiologique est très délicat à établir puisque nous ne connaissons ni la taille réelle du conduit vocal du locuteur ni son articulation spécifique. Que des normalisations nous réellement, la variation d'origine physiologique, articulatoria ou une partie des deux?

2. METHODOLOGIE

Le corpus d'évaluation est constitué de deux phrases qui comportent les réalisations des deux voyelles /e, e/, une voyelle par phrase, en syllabe accentuée et dans des contextes consonantiques symétriques : labial, dental, palatal et ralvulaire. Voici ces phrases, nous avons séparé les voyelles étudiées par un espace.

"A Papeete, son f r e r a m angé ce f e ve faite en i l g e."

"Vous ser ez racrod é si vous pou v e z pa y e r c hae mois?"

/e / e / e / e / apparaissent dans des contextes où leur prononciation est bien déterminée en français, de ce fait ces contextes ne sont pas tout-à-fait comparables : syllabe fermée pour /e/ et syllabe ouverte suivie d'une frontière morphologique pour /e/. Troize locuteurs masculins ont enregistré le corpus, que nous espérons compléter par les données d'autres locuteurs et surtout d'autres locutrices à la présentation de ce papiers au congrès. Le signal a été échantillonné à 16 KHz, et les fréquences formantiques ont été mesurées manuellement sur grand écran.

3. RESULTATS

Signalons d'abord que l'opposition /e, e/ est respectée puisque les réalisations acoustiques de ces deux voyelles sont bien différentes dans le contexte donné. D'autre part le facteur de chaque contexte consonantique est parfaitement prévisible que si le locuteur.

Nous avons testé deux procédures de normalisation:

- la normalisation conjointe des locuteurs et des locutrices,
- la normalisation séparée des locuteurs et des locutrices à partir des moyennes formantiques masculines et féminines.

Voici les résultats obtenus avec la première méthode, nous avons mis entre parenthèses les résultats obtenus avec la deuxième méthode quand ils sont différents des précédents. Trois formants ont été utilisés pour l'apprentissage :

- 22% (30%), 27% (33%), 76% (76%) pour F1, F2, F3 de /e/,
- 5%, 61%, 66% pour F1, F2, F3 de /e/. Nous ne constatons pas de grandes différences entre les méthodes qui normalisent les voyelles des hommes et les femmes ensemble ou séparément, à une exception près: le premier formant de /e/ en contexte palaté, mieux normalisé par la deuxième méthode ; pour /e/, les droites "physiologiques" sont sensiblement identiques pour les deux méthodes.

Il sera intéressant de vérifier ces résultats sur les formants des voyelles d'arrière ouvertes. L'apprentissage effectué avec la voyelle /e/ se révèle aussi performant pour la normalisation de /e/ que l'apprentissage effectué avec /e/ même. Là encore, il sera intéressant de confirmer ces résultats avec d'autres voyelles.

Les résultats obtenus par voyelle et par formant semblent satisfaisants si on considère que seules des voyelles émises par les locuteurs masculins ont été réellement normalisées. L'enquête à barks à la place des Hertz amémore légèrement les résultats obtenus pour F1 en équilibrant pour chaque formant les écarts fréquentiels entre les moyennes ou références acoustiques.

L'articulation spécifique à chaque locuteur tourne avec le contexte et s'exprime le long de nos droites "physiologiques" ce qui perturbe les résultats obtenus pour F2 /e/. Il est difficile de comparer les résultats obtenus pour F1 /e/, la variation inter-locuteur étant relativement faible avant la normalisation.

4. REFERENCES


SOME SARA VOWEL INVENTORIES
AND VOWEL SYSTEM PREDICTIONS

D. I. Djarangar
Institut de la Communication Partielle
URA CNRS n°568 Grenoble France

ABSTRACT
Formant measurements of six Sara languages confirm the general principle that vowels of a given system tend to be sufficiently dispersed and that systems are more filled on the front and/or back ranks than on the high one. But, in vowel system predictions, Sara languages which have at least 6 vowels do not always attest a phonemic [e] which, according to Crothers [1], should appear “earlier” than phonemic [i] and [a]. Still, in Sara languages, [e] does tend to appear at the phonetic level.

The aim of this study is to confront some vowel inventories with typological [1, 4] and computational [2, 3] system predictions.

1. SARA VOWEL SEGMENTS
Sara (Central Sudanic family [5]) is a group of about twelve languages spoken in Chad and Central Africa. Formant measurements of vowel systems of six of these languages (Nar, spoken in Doro; Sar, spoken in Douyou; Mbay, spoken in Moissala; Kaba, spoken in Paoua; and two varieties of Bedjond, spoken in Bediondo and Beda) provided us with the acoustic spaces shown below (fig. 1-6). Each vowel was inserted in a carrier sentence in a [ ...VT ...] context. Twelve repetitions of each sentence were recorded, in a random order, by one speaker per language, in a soundproof booth. The corpus was digitized, and the vowel formants were measured (using the ICP’s digital signal editor, EDISIG). Figures 1 to 6 display those measures with 90% dispersion ellipses.

At a surface phonology level, a language like Bediondo Bedjond [6] has 11 vowels that could be all interpreted as phonemic, because they can all contrast in more or less identical contexts: [t[i] is swallowed, [t[e] he deceived him, [t[e] he married her many times, [t[a] he took it many times, [t[s] he tied up, [t[o] he blew on it, [t[u] he swallowed, [t[i] to, into, [t[e] he deceived, [t[e] he blew on him, [t[e] he tied him. But in underlying phonological representations, only [i, e, a, o, u] are phonemic. Actually, [a] and [i] are, respectively, allophones of [e] and [i]. Furthermore [e, o, a] are respectively phonotactic fusions of [a+e], [a+e] and [o+e]. In this study, such allophones or blends (empty ellipses) will be physically positioned with regard to other vowel realizations corresponding more closely to their underlying representations (filled ellipses). This analysis of Bediondo Bedjond is also valid for Beda Bedjond, Kaba, Mbay, Nar and Sar [7, 8, 9, 10].

2. VOWEL DISPERSION
According to the Theory of Adaptive Dispersion [3], vowels tend to be sufficiently distant in the so-called anthropophonic vowel triangle. At the same time, vowels of a given system tend to fill rather the back and/or front ranks than the high one.

Available phonological descriptions of Sara languages show that the Bedjond dialects, Kaba, Mbay, Nar and Sar all display an unbalanced phonological inventory: [i, e, a, o, u]. But allophones and phonotactic blends tend to fill gaps and make the systems more balanced, except for Sar and Mbay. As a result, the system of Nar (fig. 3) is one of the most classically balanced.

Beda Bedjond presents a very centralized [e]; in fact, it is phonetically an [a]. In Kaba (fig. 4), the central high vowels [i] and [a] drift toward their front counterparts [i] and [e] without protruding the lips. (Note that generally, our Kaba speaker tends to have a narrow front-back space vs. one of the largest high-low dimension). From a general point of view, vowels of the same aperture degree are relatively well separated. The exceptions are [a] and [a] in Bediondo Bedjond, which are two very close allophones of different phonological vowels (cf. [8] for a study of this special case, [a] like [æ], being clearly rounded).

3. VOWEL INVENTORIES AND THE LACK OF PHONEMIC [e]
According to [1], languages with five or more vowels have phonemic [e]. Most of the time, the five vowels are [i, e, a, o, u]. The prediction of [e] as the sixth vowel would require filling the front rank before the back one. The case seems to be reversed in Sara.

At the phonetic level, Sar has 7 vowels [i, e, a, o, u, i]. According to the descriptions of [8] and [10], [i] is an allophonic variant of [i]. It is interesting to notice that while Sar attests [i], it does not have [e], although, from a typological point of view, one could expect an [e] first. The acoustical vocalic triangle clearly shows here the gap left by this absence of [e] (fig. 1).

Mbay has 8 vowels: 6 peripheral [i, e, a, o, u] and 2 interior [i, a] [9]. Mbay does not attest [e]. The gap left by its absence appears clearly in the acoustic space (fig. 2).

Nar also has 8 vowels. It attests [e] but only at a phonetic level: [e] is a result of [e] in a preconsonantal position (fig. 3).

In Bedjond dialects [6], [æ] (always realized long) is a combination of a final [...]φ with the pronoun [æ] "his" (e.g. [t[i] mouth + [e] his + [t[æ] his mouth]. Only Kaba seems to have an [e] which is not an allophone, from a surface phonology point of view. But the lack of analyses available for Kaba prevents us from giving firm phonological conclusions on this system.

4. THE ORDER OF APPEARANCE OF INTERIOR VOWELS IN SARA LANGUAGES
According to [1], six-vowel systems have [i, e, a, o, u] or [i, e, a, o, i]. The second system is said to be the most common. Generally, when languages have 9 vowels, there are 7 peripheral vowels [i, e, a, o, u, o, i]. The other 2 vowels are the front rounded [y, ø], the back unrounded [u, v], or central [i, a].

Sara languages do not have 9 phonemes, but they can display 9 phonetic vowels. For Sara dialects like Beda Bedjond and Kaba, that have already filled their peripheral positions with [e], it seems easier to develop an [a] after an [i] (as in all other Sara languages) rather than developing a completely different range of vowels like front rounded or back unrounded.

Bediondo Bedjond speakers realize 11
vowels (Fig. 6). Thus, after using the central rank, Sara languages exploit the front rounded rank of vowels [o, e].

These vocalic systems allow us to suppose that when Sara dialects need to develop new vowels, beyond those cardinal and phonemic vowels represented in the following figure by Ω.

the tendency is first to develop allophones at the high central position, Ω (Sar); before filling the remaining cardinal (Nar and Beda Bedjond) or central (Mbay) positions, Ω (both, for Kaba); then, to exploit the front rounded positions, Ω (Bedjond Bedjond).

5. SARA LANGUAGES AND VOWEL SYSTEM PREDICTIONS
Again according to [1], "Languages with six or more vowels have a and also either i or e, generally the former" and "Languages with seven or more vowels have e, o or i, e;" Sara languages have 6 phonemic vowels. In accordance with Crothers' prediction, they have [a] and [e], but they do not have phonemic [i] and [a]. But at a phonetic level, all of them have [i], and 4 of them also have [a] (including [e] = [a] of Bedjond Bedjond).

6. CONCLUSION
In regard to our formant measurements and to phonological analysis of Sara languages, one can say that while Sara vocalic inventories appear to maintain a sufficient distance between vowels in a given system, they show, typologically, that languages with five or more vowels do not obligatorily tend to give rise to phonemic, or even surface, front rank filling with [e].

ACKNOWLEDGEMENTS
Many thanks to C. Abry for his fruitful suggestions and T. Sawallis for improving our English.

REFERENCES
LOCUS-NUCLEUS RELATION AND VOT IN SPONTANEOUS
AND ELICITED SPEECH

Diana Krull

Institute of Linguistics, Stockholm University, S-106 91 Stockholm, Sweden

ABSTRACT

CV-sequences occurring in the spontaneous speech of five Swedes were compared to the same sequences reproduced by each speaker in citation form words. Two acoustic characteristics are reported here: F2 trajectories for C=/b, d, m, n, /l/ and voice onset time (VOT) for C=/p, t, k/. Plotting F2 at the CV boundary (locus) as a function of F2 within the vowel (nucleus) resulted in steeper regression lines for spontaneous speech than for citation speech which was interpreted as a greater extent of contextual assimilation. Both locus and nucleus had also more central values in spontaneous speech. For VOT, no significant difference was found between spontaneous speech and citation form words, although both the duration of the consonantal closure and the following vowel were considerably shorter in spontaneous speech.

1. INTRODUCTION

The present paper reports two studies comparing CV-sequences in spontaneous speech and in citation form words. The studies are a part of a larger investigation of phonetic variation in spontaneous Swedish.

The first study deals with the extent of contextual assimilation between the consonant and the vowel; the second addresses itself to comparisons of VOT.

1.1 Locus-nucleus relation of F2 and contextual assimilation

The size of the formant excursion from the CV boundary towards its "target" within the vowel has been shown to depend to a large extent on the duration of the vowel [7,4]. Moreover, the dimension of more or less clear pronunciation - "hypo" or "hyper" speech is important: formant excursions have been shown to be larger in clear speech compared to neutral speech [10].

Plotting the locus frequency of a formant, e.g. F2, as a function of its frequency within the vowel results in a linear function called the "locus equation", Eq. (1):

\[ F_2 = kF_2 + c \]

where \( F_2 \) denotes the initial locus of the second formant, \( F_2 \) is the maximum or minimum within the vowel, and \( k \) and \( c \) are constants. Locus equations were first used by Lindblom [7] who demonstrated that the value of the constant \( k \), i.e. the slope of the regression line, varies with consonant place of articulation. The slope also expresses the extent of contextual assimilation between the consonant and the vowel [5]. In the case of maximal assimilation, \( F_2 \) at the initial locus has the same frequency as in the middle of the vowel: \( k = 1 \), and \( c = 0 \). In the other extreme, the initial locus is invariant through all vowel contexts, \( k = 0 \) and \( y \) has a constant value.

1.2 Experiment I

Five male speakers of Standard Central Swedish served as subjects. Natural continuous speech was obtained by asking each subject to relate a previously read short story, and to answering questions posed about the subject's work, travel, etc. The sessions were recorded and transcribed. Thereafter, word-initial CV-sequences with \( C=/b, d, m, n, /l/ \) and voice onset time (VOT) for \( C=/p, t, k/ \) were measured on wide band spectrograms at two points: "locus" at the CV boundary and "nucleus" at the maximum or minimum point within the vowel. If there was no minimum or maximum, the corresponding measurement was performed in the middle of the vowel. The words containing the CV-sequences used for measurement were then assembled in lists, separate for each speaker, who read the words separating them with pauses.

Plotting the locus as a function of nucleus resulted in slopes and y-intercepts given in Table I. It can be seen that, for a given place of articulation, the slope of the regression line is steeper for spontaneous speech, which can be interpreted as a sign of greater contextual assimilation. Of the possible factors influencing the extent of assimilation, the roles of lexical stress and phonological length were investigated, using only content words. The results showed that while there was relatively little change in slope with different phonological length, lexical stress caused a marked flattening of the slope. Higher k-values indicate that F2 locus and nucleus were nearer each other in spontaneous speech. However, further investigation showed that the locus and nucleus frequencies of the citation form words had not changed in a direction towards each other in spontaneous speech. Instead, both had moved towards a more central frequency value.

1.3 Discussion I

According to our interpretation of k-values in locus equations, there was always more contextual assimilation in spontaneous speech. Similar results have been obtained for French [1,2], Spanish and Catalan [11]. Both locus and nucleus frequencies were also more centralized in spontaneous speech. One probable reason for these differences is the usually shorter duration of the sequences in question and a resulting formant undershoot, i.e. the formant has not time to come near its target value [7,4]. Another reason for the difference may lie in the "hyper"-"hypo" dimension: the citation form words were usually more clearly pronounced than their spontaneous counterparts [10].

<table>
<thead>
<tr>
<th>Speaker</th>
<th>OE</th>
<th>RL</th>
<th>JS</th>
<th>PT</th>
<th>AV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference y-intercept</td>
<td>1106</td>
<td>1193</td>
<td>1041</td>
<td>1033</td>
<td>870</td>
</tr>
<tr>
<td>slope</td>
<td>0.29</td>
<td>0.31</td>
<td>0.36</td>
<td>0.36</td>
<td>0.45</td>
</tr>
<tr>
<td>Spontaneous y-intercept</td>
<td>937</td>
<td>906</td>
<td>823</td>
<td>795</td>
<td>755</td>
</tr>
<tr>
<td>slope</td>
<td>0.32</td>
<td>0.39</td>
<td>0.44</td>
<td>0.47</td>
<td>0.51</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y-intercept</td>
<td>568</td>
<td>313</td>
<td>363</td>
<td>278</td>
<td>369</td>
</tr>
<tr>
<td>slope</td>
<td>0.58</td>
<td>0.74</td>
<td>0.70</td>
<td>0.79</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Table I. y-intercept and slope for the regression lines of initial locus vs. nucleus F2 in CV-sequences.

C=/b, d, m, n/ VOT

<table>
<thead>
<tr>
<th>Speaker</th>
<th>OE</th>
<th>RL</th>
<th>JS</th>
<th>PT</th>
<th>AV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference y-intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slope</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spontaneous y-intercept</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slope</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. VOT IN SPONTANEOUS SPEECH AND IN CITATION FORM WORDS

Lisker and Abramson [8] compared VOT - the time between the stop release and the onset of the vocal cord vibrations - in isolated English words and in read sentences, showing that for a given CV-sequence VOT was considerably longer in isolated words. The role of several contextual features on VOT was studied; three of these were shown to have no effect: initial vs. non-initial, utterance tempo, and vocalic environment. Stress, on the other hand, had a strong effect. In Swedish, VOT has been shown to increase with stress in nonsense words [6][9]. In semantically meaningful sentences, VOT can be approximately doubled with the addition of emphatic stress [3]. The aim of this investigation is to study VOT in CV sequences in lexical words occurring in spontaneous speech, and in the same words read in citation form.

2.1 Experiment II

The material consisted of two of the recordings described in section 1.1 above. This time, CV-sequences occurring in content words were located, C being a voiceless stop and V any vowel. For each CV-sequence, the duration of the stop gap, VOT, and the duration of the vowel were measured on wide band spectrograms. As in the previous experiment, word lists were prepared and read by the speakers. The effect of four of these factors of possible influence on VOT are reported in this paper: (1) Stress (main and secondary); (2) place of articulation; (3) phonological length of the vowel and consonant; (4) the physical duration of the vowel and of the stop gap.

The results of the comparison revealed no significant difference between VOT in spontaneous speech and in citation form, although VOT tended to be slightly shorter in the isolated words (Table II). There was, however, a large difference in both the duration of the stop gap and that of the vowel: both were much longer in citation form words.

Of the different factors whose influence on VOT was studied, only stress and place of articulation were shown to have a strong effect, both in spontaneous speech and in citation form words. Mean VOT was between 30% and 100% longer in stressed CVs than in corresponding unstressed syllables. In spontaneous speech as well as in citation form, the velar consonant had a longer mean VOT than the dental and labial. The mean VOT for the dental consonant was in most cases longer than that of the labial. For both overlap in VOT between places of articulation as well as stressed and unstressed syllables.

According to t-tests, neither the phonological length of the vowel nor that of the consonant had a significant effect on VOT. There was, moreover, no significant correlation between the physical duration of the vowel and VOT. On the other hand, there was a weak but significant (p<.05) negative correlation between the duration of the stop gap and VOT. (See [6] for detail).

2.2 Discussion II

Lisker and Abramson's data [8] showed considerably longer VOT for words read in isolation than for words read in sentences. It was therefore surprising to find that in the present material VOT in isolated words tended to be slightly shorter than in spontaneous speech, although the difference was not significant. The standard deviations for VOT were also approximately the same in both speaking styles, showing that the variation in VOT was not larger in spontaneous speech. The duration of the stop gap and that of the vowel, on the other hand, were both much longer in citation form words. There was also a greater variation in duration. The difference between the present results of Lisker and Abramson [8] may be due either to language differences or to the fact that the connected speech in the American material was read text, while the Swedish material consisted of spontaneous speech. Possible differences in VOT between these two speaking styles have not been investigated.

FOOTNOTE

1 The project: Speech transforms - an acoustic database and phonetic and phonological rules for Swedish phonetics and phonology (Ole Engstrand, project director, Diana Krull, Björn Lindblom and Rolf Lindgren), supported by The Bank of Sweden Tercentenary Foundation, grant 86/109 and by the Swedish Board of Technical Development, grant 89-0027.

REFERENCES


Table II. The duration of the closure (stop gap), VOT, and V2 (in ms) in spontaneous speech and in citation form words. CV-sequences in word-initial position are not included.

<table>
<thead>
<tr>
<th>Speaker JS</th>
<th>Closure SD</th>
<th>VOT SD</th>
<th>Vowel SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstressed CV</td>
<td>sp</td>
<td>77</td>
<td>26</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>ci</td>
<td>163</td>
<td>47</td>
<td>34</td>
</tr>
<tr>
<td>Stressed CV</td>
<td>sp</td>
<td>63</td>
<td>17</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>ci</td>
<td>121</td>
<td>44</td>
<td>47</td>
</tr>
<tr>
<td>Speaker PT</td>
<td>Unstressed CV</td>
<td>sp</td>
<td>90</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>ci</td>
<td>137</td>
<td>40</td>
<td>32</td>
</tr>
<tr>
<td>Stressed CV</td>
<td>sp</td>
<td>67</td>
<td>13</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>ci</td>
<td>72</td>
<td>15</td>
<td>56</td>
</tr>
</tbody>
</table>
A STUDY OF [r] AND [e] IN SPONTANEOUS SPEECH

Carme de la Mota Gorrix
Laboratori de Fonètica, Universitat Autònoma, Barcelona, Spain

ABSTRACT

This paper describes the acoustic features of the Spanish [e] and [r] both in spontaneous and in laboratory speech. The results discussed below show that a shorter duration and the reduction of the ratio values of the differences between the second formant and the following vowel are of prime importance in spontaneous speech. However, there is a close relation between the second formant of the consonant and that of the adjacent vowels both in spontaneous and in laboratory speech.

1. INTRODUCTION

One of the most common approaches in studying [e] and [r] has been the analysis of their duration, and therefore it is well known the description of the several phases (closed and open) in [r] (e.g. FANT [1]). Some works have also noted the importance of the vocalic context in formant frequencies in Spanish (QUJIS [2]). This paper suggests some relevant differences in duration and spectral structure for both [e] and [r] in two speaking styles.

2. EXPERIMENTAL PROCEDURE

Data from spontaneous speech have been obtained from an hour recording of speech obtained by asking the subject—a male Spanish speaker—about the city where he comes from, his family, his work, the military service, etc.

These data have been compared with those obtained by studying [e] and [r] in laboratory speech, which is, embedded in carrier sentences which were read at a normal speech rate by the same subject.

The registration was made in a soundproof room at the Phonetics Laboratory at the Universitat Autònoma de Barcelona, using a Revox A/77 tape recorder and a Shure 515 Sb Unidyne microphone.

The corpus was then low-pass filtered, digitized at a sample rate of 10 KHz, and stored. It was analysed by means of broadband spectrograms using a Mac Speech Lab II programme.

Both [e] and [r] have been studied in intervocalic contexts, either in stressed and unstressed syllables.

A whole of 300 items—uttered in laboratory speech—and 445 items—coming from natural speech—have been segmented and measured. A simple statistic analysis have been performed to extract mean values of the consonant duration, the four first formant frequencies of these consonant, and the four first formant frequencies of the C-V transition starting point.

Intensity values are not studied in this paper. However, it should be interesting to have also into account the strong decrease in the sound pressure level of the consonant in further research, as it has been pointed out by CHAPCOULOFF [3].

3. RESULTS

3.1. Duration

It has been pointed out that one of the most important differences between continuous speech and laboratory speech is duration. There is indeed a shortening phenomenon which is closely related to the fact that the speaking rate is usually much faster in continuous speech. Spanish [e] and [r] are shorter in spontaneous speech, as is shown in Table 1.

TABLE 1: Mean duration values (in milliseconds). Comparison between laboratory and spontaneous speech.

<table>
<thead>
<tr>
<th></th>
<th>Laboratory Speech</th>
<th>Continuous Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>31.6</td>
<td>27.1</td>
</tr>
<tr>
<td>Total</td>
<td>66.6</td>
<td>45.31</td>
</tr>
<tr>
<td>1r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 stages</td>
<td>16.9</td>
<td>17.1</td>
</tr>
<tr>
<td>5 stages</td>
<td>58.65</td>
<td>79.43</td>
</tr>
</tbody>
</table>

Some remarks can be made about [r]. Its duration depends on the number of its closed and open stages. In speech laboratory [r] can be uttered with three or five different phases, and it affects the total duration as it is pointed at Figure 1. The results obtained by means of a t-statistic test prove that there are two different populations indeed, so that the degree of significance is 0.000.

A statistical study of the several phases for each type of [r] shows that differences among them are not significant as for duration.

The mean values of [r] duration in laboratory speech, taking into account the two kind of populations are those in Table 2.

However, these three or five stages do not appear in spontaneous speech. There are at most two different phases, a closed phase—the first—and an open one, and it is interesting to mark that the open stage presents a concentration of energy in the upper zones of frequency. A t-statistic test suggest us that each phase lasts approximately the half of the whole duration. The mean values are: 24.5 ms for the closed phase and 22.18 ms for the open one.

On the other hand, as for [e], there is a significant difference in milliseconds between spontaneous speech and laboratory speech. The mean values at Table 1 show the same differences observed at Figure 2 and Figure 3.

Both [e] and [r] are shorter in continuous speech, although [e] is always the shortest one.
3.2. Formant frequencies

The mean frequency values for the four first formants are those at Tables 3 and 4. However, note that, as an hour of spontaneous speech reports ur much more cases of A-[ r ]E than of U-[ r ]U, for instance, these values have been obtained by homogenizing the number of cases with each vocalic context in spontaneous speech. Otherwise, the values are not able to be compared with those obtained in laboratory speech.

**TABLE 3:** Mean frequency values for [ε] in laboratory and spontaneous speech. (Hz.)

<table>
<thead>
<tr>
<th></th>
<th>Laboratory speech</th>
<th>Spontaneous speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>F4</td>
<td>3412.23</td>
<td>3489.11</td>
</tr>
<tr>
<td>F3</td>
<td>2304.06</td>
<td>2287.24</td>
</tr>
<tr>
<td>F2</td>
<td>1384.85</td>
<td>1359.29</td>
</tr>
<tr>
<td>F1</td>
<td>367.75</td>
<td>405.98</td>
</tr>
</tbody>
</table>

The differences between laboratory and continuous speech in the steady state of the consonant do not seem to be very large. Furthermore, the consonant shows the same behaviour in both cases: the first and the second formant depend on the vocalic context, as it is shown in tables 5 and 6.

**TABLE 4:** Mean frequency values for [ɛ] in laboratory and spontaneous speech. (Hz.)

<table>
<thead>
<tr>
<th></th>
<th>Laboratory speech</th>
<th>Spontaneous speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>F4</td>
<td>3597.05</td>
<td>3361.74</td>
</tr>
<tr>
<td>F3</td>
<td>2067.2</td>
<td>2023.04</td>
</tr>
<tr>
<td>F2</td>
<td>1129.48</td>
<td>1201.34</td>
</tr>
<tr>
<td>F1</td>
<td>434.4</td>
<td>413.13</td>
</tr>
</tbody>
</table>

By the other hand, some differences between spontaneous and laboratory speech can be stated as for frequencies.

The relationship between the second formant steady state of the consonant and the transition starting point depends on the following vowel, but differs because of the speech style. This relationship would be even more evident if we took into account the steady state of the vowel. Note that the difference between the two points is higher in palatal than in velar contexts. But, anyway, differences are always higher in laboratory speech than in spontaneous speech. This fact can be expressed by means of percentages, as is shown in Table 7.

**TABLE 7:** Percentages. Difference between the second formant frequencies of the steady state of the vowel and the following transition starting point. Comparison between spontaneous and laboratory speech.

<table>
<thead>
<tr>
<th></th>
<th>Laboratory speech</th>
<th>Spontaneous speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2</td>
<td>10.7%</td>
<td>2.9%</td>
</tr>
<tr>
<td></td>
<td>7.2%</td>
<td>1.6%</td>
</tr>
<tr>
<td></td>
<td>19%</td>
<td>no cases enough</td>
</tr>
</tbody>
</table>

3.3. Spectral distribution

In fact, the results suggest that spontaneous speech favours the concentration of energy in the upper zones of frequency. About a third of the studied cases of [ε] in spontaneous speech show aperiodic energy in the higher frequencies, and about the ninety per cent of cases of [ɛ] are periodic frictions. The fourth formant is the most intense in many cases. However, both [ε] and [ɛ] are completely periodic in laboratory speech.

Figures 4 and 5 show some of the spectral differences observed between continuous and laboratory speech for [ε] and [ɛ].

4. CONCLUSION

Speaking style differences are found on duration, which is shorter in spontaneous speech, and on the reduction of the ratio values of the frequency differences between the steady state of the second formant of the consonant and that of the next vowel. Further research should pay attention to intensity levels of [ε] and [ɛ] in Spanish and to their spectral distribution in spontaneous speech.

5. REFERENCES

AUTOMATIC CLASSIFICATION AND FORMANT ANALYSIS OF FINNISH VOCALS USING NEURAL NETWORKS

Toomas Altosaar and Matti Karjalainen
Helsinki University of Technology
Acoustics and Speech Processing Laboratory
Otakaari 5A
02150 Espoo, Finland

ABSTRACT
In this paper we report results from a study of using feedforward neural networks with error back-propagation in order to see their inherent ability to learn speaker independent classification and formant analysis of Finnish vowels.

1. INTRODUCTION
The recognition and analysis of vowels is an important problem in the field of speech recognition and phonetics. Neural networks [5] are shown to give excellent performance in many speech recognition subtasks [11,12]. They can be described as "black-boxes" that given an input and desired output can actually learn to associate the input with the output. The performance levels achieved with neural nets can be very high and their use is an attractive method when performing vowel recognition or analysis [5].

In our study we used feedforward nets with error back-propagation. Figure 1 shows a possible network topology (general structure of a feed-forward set).

2. VOWEL RECOGNITION
For our vowel recognition experiments we used speech taken from 12 female and 24 male speakers. Static auditory spectra (256 in total) each consisting of a 48 point real-valued vector were used as the input representation [2]. The topmost curve in figure 2 shows the auditory spectrum of the vowel /a/. The 0-24 Bark critical-band scale corresponds to approximately 0.15 kHz.

We defined a criterion for when a neural net had learned all of the input material: a) all of the inputs had to be correctly classified, and b) a 0.75 minimum level had to be reached for the correct output layer node. The target values during training were 0.0 or 1.0.

In the first experiment we determined how many nodes were required in the hidden layer as well as which spectral representation performed best to correctly learn 8 vowels from a single male speaker. What is meant by spectral representation is the scale or resolution of the input data. We used a Gaussian band-pass filter to the original auditory spectra to obtain a fine-scale representation that would emphasize formant-like local structures in the spectrum. A higher level of smoothing was also applied to yield a coarse-scale representation that emphasized more global spectral trends. The fine and coarse representations for the vowel /a/ can also be seen in figure 2.

We then trained 100 separate nets with similar initial parameters of dimension 48-3-8 (48 input nodes, 3 hidden nodes, and 8 output nodes each corresponding to one of the eight Finnish vowels). We repeated this test for 4 to 9 hidden nodes, and for all three representations. The results which can be seen in figure 3 indicate that the fine spectral representation learned the 8 vowels most frequently, followed by the original and coarse representations. This result is explainable since emphasized formants help to distinguish each of the eight vowels of a single speaker.

For a larger input set (24 male speakers, 192 vowel spectra) these results changed somewhat and are shown in figure 4. Here the number of nodes was varied between 3 and 14 and only the original and fine spectral representations were compared. The ability of learning the input set perfectly when using the fine resolution was always lower than for the original representation. A possible explanation for this is that in general the fine representation will emphasize formants, and since several examples of each vowel exist in the training set with different formant frequencies, the variability of the input representation increases making it more difficult for the net to learn the differences. For this reason we decided to use only the original spectral representation in the remaining tests.

For all three sets the number of hidden nodes was varied from 3 to 48. Figure 5 shows the learning ability for the 24 male set. Each test was repeated 100 times to gain statistical confidence. With eight hidden nodes approximately 80% of the nets were able to learn the male training set entirely. No significant difference in performance level was observed if 10 nodes were included or not. This result is somewhat surprising because it is often assumed that human listeners do spectral normalization based on the pitch of the speaker.

For the female and male+female training sets the results were similar to the male

Figure 1. A possible network topology (general structure of a feed-forward set).

Figure 2. Original, fine, and coarse auditory spectrum representations of /a/.

Figure 3. 48-X-8 Net's Ability to Learn 8 Vowel Spectra

Figure 4. 48-X-8 net's ability to learn 192 male spectrums.
training set test, i.e. no significant improvement or degradation of learning frequency was found by including pitch information.

3. FORMANT ANALYSIS
The second main topic of this study was to investigate the usefulness of neural networks in analyzing continuous parameters or features of vowels. Specifically we wished to teach nets to be able to identify the location of the first two formant frequencies of vowels in the auditory spectrum. A traditional method to perform this task automatically is to calculate the envelope of the spectrum and peak-pick the formants. Another method utilizes solving for the poles by LPC.

We trained networks of dimensions 48-X-2, X ∈ [2,15] to estimate the two first formant frequencies F1 and F2 of vowels. These estimates were based on the auditory spectrum input and we hypothesized that the network could be more robust to the traditional methods to find and label the formant frequencies. The output level nodes of the net were modified by removing the sigmoid non-linearity thus allowing continuous valued output values to be realized. As a training set we selected 64 vowels and diphthongs from a single male speaker. The formant frequencies were labeled by hand by an experienced speech scientist.

Figure 6 shows the average F1 and F2 absolute errors as a function of the number of hidden nodes. F2 exhibits a larger error since a larger input variation exists for it but drops down to 0.15 Bark when the number of hidden nodes is seven or higher. This error corresponds to approximately 35 Hz at 1.5 kHz. The F1 error being considerably smaller was found to be 0.08 Bark which corresponds to 10 Hz at 400 Hz.

Figure 5. 24 male speakers with and without pitch information.

Figure 6. Average Formant Analysis Error of 64 Male Spectra as a Function of Net Size.

We evaluated the performance of the 48-12/2 net on three independent (with respect to the training set) evaluation sets: male (3 speakers), female (3 speakers), and male+female (3 male and 3 female speakers). As can be seen in Figure 7 the average absolute error for F1 (labelled "F1 error") when evaluated on the male set of spectra (3M) was 0.5 Bark, and for F2 (labelled "F2 error") 0.8 Bark. The F2 error was very large when evaluated on the female set (3F) - 2.2 Barks which corresponds to ~600 Hz at 1.5 kHz. Notice that the net was trained by data from a single male speaker.

To see if we could reduce the average absolute F2 error for females we trained a similar net with the original 64 vowels and diphthongs but also included eight static vowels from one female speaker. When re-evaluated on the independent sets the F2 error (labelled "F2 error 1F") as seen in Figure 7 was substantially smaller dropping to ~1.3 Barks which corresponds to ~330 Hz at 1.5 kHz for the female (3F) evaluation set.

The overall accuracy for the formant analysis tests was not always good but the nets showed a robust behavior avoiding gross errors such as incorrect formant ordering, which is very difficult by traditional methods. We also observed that networks based the formant estimates on the general shape of the auditory spectrum but did not generalize to search for exact auditory peaks. Further studies are needed to see how accurate and robust the method could be if a more complex net is used with more training material.

4. COMPUTATIONAL ENVIRONMENT
These experiments were carried out on an object-oriented signal processing environment called QuickSig [3], developed in our laboratory. QuickSig, which is an extension to the Symbolics Common Lisp and Flavors environment runs on Symbolics Lisp Machines. To speed up the tests by a factor of 150 over the Symbolics Lisp Machines a Texas Instruments TMS320C30 digital signal processor was used.

Figure 7. Evaluations of Trained Net on Independent Spectra.

5. SUMMARY
This study has shown that neural networks are very useful tools in the classification and analysis of vowels. The ability of a neural network to generalize is an attractive feature since this means that a trained net, even if it has not seen a certain input before, can make an intelligent decision.

Specifically we found that F0 does not help in achieving better performance levels for vowel recognition. This confirms earlier work [4]. The number of nodes in the hidden layer was found to affect the learning potential. With too many nodes the net will learn but will not generalize (it will learn each training element individually). On the other hand, given too few nodes all the inputs will not be classified correctly. We also found that the preferred spectral representation when having to choose from a set of representations derived from the auditory spectrum was the unmodified auditory spectrum itself.

In the formant frequency analysis experiments more spectra need to be used to verify the accuracy and potential of the approach. Eventhough performance may not reach the levels of other well established methods such as LPC, neural networks may provide a useful general indication of formant locations for later, more detailed analysis, or rule-based combination of multiple methods.

6. REFERENCES
BULGARIAN VOWEL CLUSTERS AND STATISTICS
By 30 MALE AND 30 FEMALE SPEAKERS

Philip Christov
Private consultant,
P. O. Box 895, 1000 Sofia, BULGARIA

ABSTRACT
In this paper the output is presented of the computer aided analysis of the Bulgar-ian vowels in /b-b/ context uttered by 30 male and 30 female professional speakers in stressed and in unstressed position, namely the print out of the populations (or vowel clusters) in the F1 vs. F2 plane of equally labeled vowel utterances with their cluster statistics; means, standard deviations, maximum, minimums, skewnesses and kurtoses.

1. INTRODUCTION
In a previous paper [1] an algorithm has been reported which makes use of phonetic knowledge to perform computer aided analysis of speech followed by formant tracking. It has been described lately how this algorithm has been applied to the processing of a phonetic material [2] from the Bulgarian Central Allophones Data Base [3].

Here will be presented in more details the direct output of the computer processing of this phonetic material.

2. VOWEL CLUSTERS
Central (5-Vowel-b) allophones of the vowels: /i/, /e/, /a/, /e/, /o/
are uttered in Standard Bulgarian by 30 male and 30 female professional speakers in stressed and in unstressed position. The allophones are imbeded in words (See APPENDIX) uttered with falling intonation at the end of a standard carrier sentence. The labeled sound recordings of the vowel utterances are verified by a group of 20 listeners so that the incorrect utterances to be rejected by the computer and only the correct ones to be admitted to further processing.

The analog speech signal is digitalized with a sampling frequency of 20 kHz and then processed with an IBM 360/40 computer.

The computer performs a FORTRAN program based on the subroutine PORTF from the SSP [4] and built up according to the algorithm [1]. The computer produces, except of output listings of the labeled points (F0, F1, F2, F3) ....... /w/ where

LABEL = /phonic symbol
/presence or absence of stress/sex of speaker/
also two dimensional plots of the sets of equally labeled points in the space of the first two vowel formants (See Fig. 1 to 4).

It can be seen that the number of points in the clusters on the plots is sometimes smaller than the number of the speakers in each group. This effect is obtained because of:
1) The incorrect utterances rejected by the group of listeners;
2) The coinciding points in the F1 vs. F2 computer print out;
3) Some single points very distant from the clusters nuclei which got out of the F1 vs. F2 computer print out.

There are in fact only three such points exclusively in the female utterances, namely two points in the /i/ cluster above the upper limit of the graph and one more in the cluster of the vowel /u/.

The number of coinciding and out-of-the-graph points is presented in the last column of Tab. 1 to 4.

3. CLUSTER STATISTICS
The statistical processing of all vowel utterances verified by the listeners is performed by a FORTRAN program presented in this paper. The use of the SSP subroutine [4], among them the subroutine TALLY to compute means, standard deviations, maximum, minimums and the subroutine WOMAN to help by the computing of the skewnesses and kurtoses. The statistical estimates, computed for each cluster of equally labeled points, are presented in Tab. 1 to 4. In the bottom part of each table are presented the statistics of the overall population of the six vowels above.

4. DISCUSSION
As the behavior of the vowel clusters in dependence of the sex of the speakers and of the kind of uttering them is discussed elsewhere [2] it will be only mentioned now that the results of the computer processing of the raw experimental material reported here support the inferences deduced from the sets of manually determined closed loops in [2].

5. CONCLUSION
The phonetic data presented in this paper may be of use to the scientific community by trivial and computerized comparative phonetics studies and by machine synthesis and recognition of Bulgarian speech.

6. REFERENCES

APPENDIX:
Word list in rough phonemic IPA - transcription:
STRESSED - UNSTRESSED
/biblja/ /bibljejki/ /b'eb/ /beet ef/ /b'ab/ /bob/ /b'obrev/ /b'obrev/v'den/ /b'oba/ /bob/ /b'oba/ /b'obrev/ /b'oba/ /bob/
Fig. 1. First two formants computer graph of the six Bulgarian vowels uttered in stressed position by 30 male speakers

Fig. 2. First two formants computer graph of the six Bulgarian vowels uttered in unstressed position by 30 male speakers

Fig. 3. First two formants computer graph of the six Bulgarian vowels uttered in stressed position by 30 female speakers. There are three points /æ/ with rather high second formant which got out of this graph. Two of them are labeled as /a/', (F1=624, F2=3744) and (F1=850, F2=3025), and one as /a/, (F1=1176, F2=3864)

Fig. 4. First two formants computer graph of the six Bulgarian vowels uttered in unstressed position by 30 female speakers. The point /&/, (F1=768, F2=1536), which coincides with a point of the /a/-cluster, is not marked on the figure

TERMINOLOGY:

LEGEND TO THE FIGURES:
In the computer print out capital letters from the latine alphabet are used together with the symbol "ape" to designate some symbols of the International Phonetic Alphabet (IPA) as follows:

I = /i/: A = /a/: 
E = /e/: U = /u/: 
O = /o/: 

REMARK:
With a single exception (see text to Fig. 4) coinciding points belong to one and the same vowel cluster.

LEGEND TO THE TABLES:

n - number of vowel utterances admitted to analysis after being verified by a group of 20 listeners 
c - number of positions in the F1 vs. F2 plane in which the coordinates of each two or more vowels do coincide or a single vowel gets out of the computer print out
Turkic phonology is the phonology of synharmosism. The model of the phonology of synharmosism is proposed. The synharmonic script theory is worked out and the system of the syllabic Turkic script is proposed.

None of the accepted at different times graphs in Turkic languages - Arabic, Latin or Russian - was an optimum script from the point of view of phonological and phonetic nature of the Turkic speech. Since in the first place it was necessary to introduce quite a number of additional letters in the second place extra orthographic and orthoeptic rules were needed, in the third place the main shortcoming of these scripts was that the principle "one sound for one symbol" was adopted. While successive and systematic synharmonic consonance of syllables in Turkic speech, required syllabic principle of the script. Apparently it is not accidentally that ancient Turkic runic script was just as such.

The script must be formed on the basis of the phonological and phonetic structure of the given language (or cognate group of languages). Only in that case graph and orthography will be rational and easy for mastering this script. Synharmonism is such means for Turkic languages, and it permits to construct an easy and economical Turkic script.

Synharmonism is not an ordinary phonetic phenomenon, but a basis of the whole linguistic structure of the (as all languages) it is a specific language unit forming the integrity of syllables and words in Turkic speech.

Here is the model of synharmosism, built as a "circle" because both synharmotypes (palatal and labial) as well as all the four synharmotimbers (hard nonlabial, soft nonlabial, hard labial, soft labial) together make up the phonological system of the Kazakh language. The main thing in this model is the equal relevance of all its components.

The upper half of the circle reflects hard (complete line), the lower half - soft (dotted line), the left half - labial (chain of circles), the right half - nonlabial (absence of circles) synharmotypes. Palatal and labial synharmotypes do not function separately however. Four independent and compound synharmotimbers are formed out of their combination: hard labial (chain of circles joined by a complete line); soft labial (chain of circles joined by dotted line); hard nonlabial (complete line and absence of circles).

Here are four timbres forming the system of synharmosism. The middle circle reflects distribution of vowels in the synharmosystem. Crossed squares indicate the absent vowels in the vocalic system (in this case the Kazakh language which is one of the Turkic languages). The inner circle reflects the synharmosystem of consonants. It is open from all the sides, which indicates simultaneous presence of all the four synharmotimbers in the system of consonants. Such universality of consonants in contrast to vowels permits to use them as basic sounds in constructing the synharmonic script.

The style of formalism of the proposed model may be subjected to criticism, and we shall be glad if someone will manage to give more efficient and adequate definition of synharmosism and to construct the appropriate variant of the model. We want, to remind that nobody succeeded in constructing a good working model of synharmosism at least those referring to "harmony of vowels". That is why it is necessary to seek and to seek. In order to succeed it is necessary to have a strict synharmonic theory, ensuring true linguistic interpretation of primordial phonetic phenomena in Turkic languages. For all this one must not be afraid of seeming or factual contradictions of this theory with established well-known theories of "eurocentrism" trend in Turkicology. It is law-governed: phonology of the language, which differs from Indo-European languages can not be explained by the theory, ensuring linguistic interpretation of phonetic phenomena in these (indo-european) languages. By the way, our knowledge of the nature and functions of synharmosism turned out to be insufficient and erroneous.

So far as synharmosism is the phonological basis of the proposed system of the script, we use the simplified term "syngramma" for designating the syllato-symbol. Graphs of syngrams are not elementary: they consist of the joining of only straight lines (we intentionally avoided rounded, oval, curved and other complex lines for the scripts) and there are only three of them. Each syngram consists of the combination of the three straight lines: vertical line "|" which is basic for all syngrams; horizontal "-" place, number and direction of its joining with the basic line indicates the type of the consonant; oblique "\"
place, number and direction of its joining with the basic line indicates the synharmonic timbre of the syllable and the phonological type of the vowel.

Syngrammas are constructed according to certain logical principle (the basic being articulatory and acoustic features of sounds) which facilitates mastering the script. This principle helps to manage with minimum of rules and exceptions to them (unfortunately, we can not give here a detailed and accessible description of the rules of the script, because of the limited number of sheets and we limit ourselves to the illustration of Consonant Symbols by Syngrammas of the consonant [p] and examples of their linear sequence.

THE MODEL OF THE PHONOLOGY OF SYNHARMONISM

NOTE:

CONSONANTS

\[
\begin{array}{c|c|c|c}
\text{CONS} & \text{SILLABE MARKED} \\
\hline
p & b & m & \text{ap} \\
\hline
t & d & n & \text{ap} \\
q & g & \text{ap} \\
k & z & \text{ap} \\
\hline
s & z & \text{ap} \\
\hline
j & \text{ap} \\
\hline
w & \text{ap} \\
\hline
\end{array}
\]

VOWELS

open

close

diphthong

for example:

SINHARMOTEMIRS

non-palatalised

palatalised

labial

\[
\begin{array}{c|c}
\text{CONS} & \text{SILLABE MARKED} \\
\hline
p & \text{ap} \\
\hline
t & \text{ap} \\
q & \text{ap} \\
k & \text{ap} \\
\hline
s & \text{ap} \\
\hline
j & \text{ap} \\
\hline
w & \text{ap} \\
\hline
\end{array}
\]

\text{türük tilderi:}

qazaq tuba

\text{non-palatalised}

\text{palatalised}

\text{labial}
DE L'INDÉPENDANCE DU PHONÈME FAIBLE AU SYSTÈME PHONOLIGIQUE DE LA LANGUE RUSSSE
S.N.Dmitrenko

Institut de la langue russe, Moscou, URSS

ABSTRACT
In present article the question is about consequences which follow from an adoption of the thesis about independent phonological status of a weak phônome in the phonological system of Russian language and possibilities of construction such phonological model in which two independent phonological unit present.

1. La reconnaissance du phonème faible en qualité de l'unité phonétique indépendante est contraire à la thèse générale de l'Ecole phonologique de Moscou à sa conception traditionnelle [1] comme à la conception présentée dans le livre d'Avanesov R.I. [2] à savoir à la thèse de la connexion de la phonologie avec la morphologie, du phonème avec le morphème de la série de phonèmes avec l'identité des morphèmes. L'adoption de la thèse du statut indépendant phonologique du phonème faible au système phonologique de la langue russe entraîne une série de conséquences dont, à notre avis, il faut tenir compte en dérivant le système phonologique de la langue russe.

1.1. Le renoncement à utiliser la série de phonèmes comme liaison entre la phonologie et la morphologie, la série de phonèmes d'Avanesov R.I. [2] (ici il s'agit de la série de phonèmes dirigée par le phonème fort), n'est pas universelle c'est-à-dire n'embrasse pas la majorité des cas. Le plus souvent le phonème faible est présenté dans la situation d'hyperphonème c'est-à-dire dans la situation qui ne peut pas être vérifiée par la position forte.

Par ex., dans la combinaison des phonèmes consonnes il y a approximativement deux fois plus de combinaisons avec des phonèmes faibles dans la position initiale que de combinaisons avec des phonèmes forts [3].

1.2. L'inclusion du phonème faible comme unité phonétique ayant le statut phonologique indépendant et la fonction de distinction sémantique (et conformément ayant le rendement fonctionnel) dans la composition des phonèmes de la langue; ainsi la composition des phonèmes c'est la composition des phonèmes forts et faibles (37 phonèmes consonnes forts, 15 phonèmes consonnes faibles de dureté-molesse, 12 phonèmes consonnes faibles de sourdité-sonorité, 5 phonèmes consonnes faibles de dureté-molesse et de sourdité-sonorité; 5 phonèmes voyelles fort et 2 phonèmes voyelles faibles /a/ et /i/.

1.3. La reconnaissance de l'existence dans les positions fixées des phonèmes faibles de tel ou tel signe et simultanément de la non-existence des ces phonèmes dans les mêmes positions comme des phonèmes forts d'autre signe. Par ex., si dans le livre d'Avanesov R.I. [2] dans la position de fin du mot sont présentés des phonèmes faibles de sourdité-sonorité et simultanément des phonèmes forts de dureté-molesse, c'est-à-dire la même unité phonétique peut être le phonème faible de tel ou tel signe et fort d'un autre signe, tandis que nous présentons des phonèmes consonnes faibles de sourdité-sonorité qui sont dures ou moux. Par ex., put., phon. /pət/, /t/ - chez Avanesov R.I., est le phonème faible de sourdité-sonorité mais fort de dureté-molesse et chez nous - /t/ - le phonème faible de sourdité-sonorité qui peut être opposé l'autre phonème faible de sourdité-sonorité, par ex., dans la forme du mot put de puty où le phonème faible de sourdité-sonorité /t/ est présentée.

1.4. En déterminant des positions concrètes de la distinction maximum et minimum il faut avoir en vue que la même position la même unité phonétique ne peut pas être présente...
comme le phonème faible de
tel ou tel signe et comme
le phonème fort d'autre
signe. Par ex., nous consi-
dérons la position de la
distinction maximale pour
les phonèmes forts comme la
position précédant les voy-
elles excepté /e/ à la li-
mite du thème de la flexion
ou sont présents les pho-
nèmes consonnants faibles de
dureté-molesse (dans cette
position la distinction des
consonnants de la dureté
et de la molesse est absente).
Devant les voyelles tous
les phonèmes consonnants forts
sont opposés (par ex., /p/ /b/
/g/ /k/ /s/ de face, /p/ /b/
de père et père - /b/ /g/ /m/ /gr/ /v/ non etc.)
Dans les autres positions
(devant /e/ à la limite du
thème et de la flexion à la
fin du mot et aussi devant
les consonnants) se présen-
tent les phonèmes forts
(surtout non-appariés de
tel ou tel signe) ainsi que
(principalement) les pho-
nèmes consonnants faibles.
Nous distinguons la distribu-
tion nette dans les posi-
tions: devant /e/ à la li-
mite du thème et de la flexion
- les phonèmes faibles de
dureté-molesse (par ex.,
/na/ /to/ /1, /e/ /o/ /s/ /f/ /e/,
/na/ /k/ /r/ /e/), à la fin du
mot - non-apparié /c/, /k'/
/h/ les sonores dures-
molles et /j/ et les pho-
nèmes consonnants faibles de
sourdité-sonorité; devant
les consonnants se présentent
surtout les phonèmes
faibles (de dureté-molesse,
de sourdité-sonorité et des
phonèmes faibles de deux
signes) et aussi les pho-
nèmes consonnants forts non-
appariés et devant /m/,
outre cela - /t/, /t'/, /d'/
/ř/, /a/, /a'/, /a/ /a'/.

1.5. La reconnaissance
du phonème faible en qualités
de l'unité phonétique
indépendante, et la renon-
cement à l'emploi de la
série de phonèmes entraine
le renoncement de la tra-
scription morphophonémate-
tique [2, § 77, p. 221-224].
Nous proposons d'employer
seulement la transcription
phonologique c'est-à-dire
la transcription qui pré-
sente l'aspect phonétique
de la forme du mot. Elle
présente aussi la composi-
tion phonétique du mor-
phème dans les limites de
la forme du mot.

2. La reconnaissance
du phonème faible en qualité
de l'unité phonétique ayant
le statut phonologique indé-
pendant, et l'examen du pho-
nème faible hors de la sé-
rie des phonèmes permettent
d'étudier du point de vue
de la combinaison et du ren-
dement fonctionnel une
grand couche de lexique
russe où sont présentés
tous les phonèmes faibles
vérifiés ou non par la posi-
tion forte. On peut donner
le tableau complet des pos-
sibilités combinatoires et
fonctionnelles. Les pho-
nèmes faibles aussi que les
phonèmes forts sont le com-
posant indépendant des formes
du mot. Par ex., /p/ /o/ /v/,
/a/ /t/, /ř/, /a'/ /a'/ etc. Ils possèdent comme
les phonèmes forts la fon-
tion de distinction sémanti-
tique et le rendement fonc-
tionnel qui dépend comme
dans le cas des phonèmes
forts non seulement de la
qualité du phonème même
mais aussi la place que ce
phonème occupe dans la
forme du mot par rapport
à sa structure morphosyntaxi-
te.

Par ex., les phonèmes conso-
nants faibles de dureté-mo-
lesse /p/, /b/, /t/, /d'/,
/a', /g'/, /k'/, /g'/ possèdent
le rendement fonctionnel,
les autres dans cette posi-
tion, ont le rendement fonc-
tionnel relâché; dans la
position à la limite du
thème et de la flexion tous
les phonèmes faibles de du-
reté-molesse ont le ren-
dement fonctionnel relâché.

Les phonèmes consonnants
faibles de dureté-sonori-
té /t'/, /d'/, /k'/ possède-
dent aussi le rendement
fonctionnel relâché dans
cette position. Tous les
phonèmes consonnants faibles
de deux signes ont le ren-
dement fonctionnel relâché
dans la position à l'inté-
rieur du morphème. Ainsi
la reconnaissance du pho-
nème faible en qualité de
l'unité phonétique indé-
pendante permet de con-
struire le modèle phonolo-
gique où les phonèmes forts
et faibles seront présen-
tés comme unités phoné-
tiques indépendantes.

REFERENCES

[1] АБАНЕСОВ Р.И., СКУРОВ
В.Н. (1945), "Очерк граммати-
ки русского литературного

"Фонетика современного ли-
тературного языка", М.

"Фонемы русского языка. Их
сочетаемость и функциональ-
ная нагрузка". Автореф.
докт. дисс. М.
ABSTRACT

In the past two decades, several models have been proposed in the literature aiming at the phonetic description of vowel systems. These models are based on principles using constraints from vowel production (‘articulatory ease’) and/or vowel perception (‘perceptual contrast’). In this presentation, we will discuss these theories and will attempt to relate their phonetic bases to more linguistic attributes of vowels.

1. INTRODUCTION

Speech serves as one of the most important means of communication between humans. In an accurate regulatory system, the subglottal air pressure and, at the same time, manipulation of the glottal and vocal tract muscles. Phonemes as vowels and consonants act as linguistic (phonological) units in a language, but at the same time, the corresponding articulation is subject to articulatory and perceptual demands. In phoneme models, the collection of consonants and vowels in a language is assumed to meet rules with respect to articulatory ease, and perceptual contrast and salience. We present an outline of the theories aiming at a structural description of vowel systems in relation to articulatory models. We will focus on two aspects of system structure: the internal structure, viz. the number of vowels and where vowels are positioned in the vowel space, and the external structure, apparent in the boundary of the vowel space. Further, we pay attention to how phonological demands on vowel systems can be incorporated in sophisticated vocal models.

2. INTERNAL STRUCTURE

Vowels in language principally serve a linguistic goal. Their existence helps to distinguish words and sentences, which is the case in minimal word pairs. Historically, phonology and dialectology show that vowel systems must be considered as systems which are continuously in development, rather than as collections of vowels which are fixed once and for all. Vowels may change e.g. due to accent shifts or Umlaut-effects (as e.g. in German language), to whims of fashion (some cases of disharmonization), or to some articulatory (vowel reduction). A shift of one particular vowel may induce the shift of many vowels in the system (e.g. the Great Vowel Shift in Middle English).

From a phonological point of view, the static structure of vowel systems is related to the presence of features with an articulatory basis, such as [round], [front], [high]. Every vowel is coded by its specific feature values, and the structure of vowel sets can be represented by ‘algebraic’ manipulation on the set of feasible feature value combinations.

Phonetically, system dynamics can be modeled by repelling forces between vowels (yielding push chains) or by the tendency to fill system gaps (drag chains). These effects can be understood by assuming principles of ‘sufficient perceptual contrast’ or ‘optimal contrast’, respectively (Dieter, 1983).

In vowel models, the actual linguistic vowel systems are assumed to optimize perceptual contrast and, in an extension, articulatory ease. Liljencreutz & Lindblom (1972) and ten Bosch (1986) are the first to implement a principle of optimal perceptual contrast in a so-called vowel dispersion model. In a 2D formant space, the number of vowels is such that the system contrast was maximized, by the minimization of the system quality:

\[ Q = \sum_{i < j} d(v_i, v_j) \]

where \( d(v_i, v_j) \) denotes the (Euclidean) distance between any two vowels \( v_i \) and \( v_j \) in the ‘push’ model, \( q \) is the sum taken over all distinct vowel pairs \( 1 < j \leq N \). The particular implementation chosen by ten Bosch (1986) was to add the back to generate too many high vowels for large \( N \), due to a too large perceptual distance between \( /a/ \) and \( /u/ \).

One of the basic ingredients in this approach, viz. the perceptual distance between two vowels, has later been modified to more sophisticated models for the auditory spectrum (Blaxson & Lindblom, 1981; Lindblom, 1986).

In the literature, the 2D inter-vowel perceptual contrast has been subject to further refinement and extension to 3D. The extension to higher-dimensional formant spaces is considered in Schwartz et al. (1989) and Ten Bosch (1991). These studies show the great tendency of the resulting model systems on variations in parameters controlling the perceptual distances between vowels. The best perceptual metric for nearby vowels has recently been reported to be the 2D Euclidean metric after back transformation of \( F_1 \) and \( F_2 \) (Kwisty-Port & Atal, 1989).

Since their structure, which were determined by two parameters only, this result must be carefully interpreted, leaving aside the question about the relation between the phonetic distance (that we search) and the phonetic distance (that they measure).

While it has been subject to continuous refinement, the system contrast \( Q \), however, hasn’t grown beyond the form

\[ Q = \sum_{i < j} d(v_i, v_j)^2 \]

\[ d \] now involving combinations of transformed formant frequencies (Schwartz et al., 1989) or other differences (Lindblom, 1986). The problem that we want to address here is that this expression \( Q \) is in fact arbitrary, if we are suggested by repelling forces between magnetic dipoles or dipoles, but lacking, in fact, any linguistic or even physical basis. Ten Boscht et al. (1987) propose an expression

\[ Q = \prod_{i < j} (1 - \exp(-d)) \]

for the product being taken over all distinct vowel pairs, and \( a \) is some scaling parameter. \( Q \) is to be optimized. The rationale is, that the repelling force (as \( d / q(d) \) is interpreted as a probability of two vowels on a distance \( d \) not being confused. The system quality \( Q \) would denote the probability of no confusion at all between any two vowels, under the assumption of independence of the probabilities involved. This idea has also been suggested by Lindblom in 1975, however, in this approach, however, a weak argument can be detected, namely that the resulting optimal vowel configuration can (easily) be shown to depend on the exact shape of \( q(d) \) (Ten Bosch, 1991). Moreover, the probability of two vowels being confused is not based upon any linguistic consideration at all.

In Ten Bosch (1991), another expression \( Q \) has been elaborated

\[ Q = \min\{d(v_i, v_j)\} \]

i.e. the minimum over all distances between distinct vowel pairs. Three advantages can be recognized: (a) the system contrast is related to a perceptual bottleneck of the whole system rather than to global system properties: the bottleneck is then located at the location of the nearest vowels. (b) The influence of the exact shape of the relation between inter-vowel distance and inter-vowel confusion is apparent on exactly one place in the vowel system, rather than being spread out by weighting all inter-vowel distances (as done in eq. 2). (c) Any sufficiency constraint of the system contrast is directly replaceable to the minimal perceptual distance, then obtained by minimizing eq. 2 (Ten Bosch, 1991). This yields, in my opinion, a strong argument for the latter modified \( Q \) (Ockham). Property \( a \) is particularly useful in numerical simulation of push and drag chains. Ten Bosch (1991) it is attempted to explain the emergence of diphthongs as a consequence of a local high vowel density in the vowel space. Although this model fails to explain diphthongal properties in detail, gross effects, such as for diphthongs to have a relatively large trajectory, can be clearly demonstrated.

Articulatory constraints were not explicitly dealt with, although calculations were carried out in the acoustic domain. Recent implementations attempted to combine perceptual and articulatory constraints (Bonder, 1986; Ten Bosch, Bonder & Pola, 1987; Ten Bosch, 1991). Other approaches are given by Abry, Schwartz, Brillou, Bo, Perrier, Gerst (see the references) and colleagues in Groningen. Some of these developments have been used to forward an elaborated version of the Quantal Theory (cf. Stevens, 1972), in which perceptual and articulatory constraints are combined into one principle. In these recent models, either view have been adopted (leading to e.g. the notion of focal points, articulatory plates, insufficient control of outputs). More elaborated definitions of \( Q \) have been introduced (Schwartz et al., 1989).


3. EXTERNAL STRUCTURE

We mean by external structure of vowel systems the design of the space boundary in articulatory terms. It opposes the internal structure, with which we associate articulatory problems (S, a being a large positive number). The combination of $D_A$ and $Q$ was left as too many parameters were involved in the optimization sessions. The search for a balance between $D_A$ and $Q$ turned out to be a Pandora's Box. We therefore also have the notion of 'articulatory system effort' and even forget the role of consonantal context in any definition of articulatory ease.

Another important element is the refinement of the actual articulation-to-acoustic relation. The Quantal Theory (QT; Stevens, 1972, 1989) makes use of the non-uniformity of the mapping. In its pure form, QT states that the articulatory positions of which the acoustic output (to some norm) is less sensitive to articulatory deviations are favourable over other positions (articulatory plateaus). The Quantal Theory predicts, in the case of vowels, the corresponding favoured vowels to be a member of a vowel system. The presupposition of the Quantal Theory, however, still lead to discussion and have been questioned by many authors (cf. Journal of Phonetics, vol. 17), whereas the results are not confirmed, cf. e.g. Ladefoged & Lindau, 1988; Ten Bosch & Pols, 1989.) It is generally believed, however, that the speech signal inherits 'quantal' phonetic properties as a consequence of non-linearities of the articulation-acoustic mappings and probably, the categorical perception of sounds. When quantality exists, it is probably a result of close approximations of formant frequencies (Stevens, 1989; Badin & Schwartz et al., 1989; Ladefoged et al., 1988).

We briefly return to the open question of phonological enrichment of phonetic vowel models, and perhaps unavoidable, subjective inherent to phonetic models is that they cannot easily account for vowel contrast. Although linguistic oppositions are ultimately based upon phonetic contrast, there is a relation between the need of intervowel contrast and the 'lexical load' of the opposition? The relation between phonetic contrast and phonological contrast seems to be derivable directly from the statistics on lexemes in a language. In Dutch, /a/ and /æ/ have the largest (most often frequented) minimal set in common, despite their very close pair in the Dutch vowel system.

![Fig. 1. Contour lines in the $(P_r, P_a)$ plane of an effort function defined on the articulatory space. Scaling: 1 $\equiv$ 2000 Hz.](image)

the acoustic output. Accordingly, the minimum effort values define a 'effort landscape' on the acoustic space. It is shown in Ten Bosch (1991) that a relatively simple effort function $c$ can be found such that the boundary of the vowel space, as found in languages, resembles closely one of the contour lines of that landscape (fig. 1).

Acknowledgements

This study has been supported by the Dutch Organization for the Advancement of Pure Scientific Research NWO (project 300-161-030).

4. REFERENCES


ARTICULATORY AND PERCEPTIVE ASPECTS OF TYPELOGY OF SOUND SYSTEMS IN CONDITIONS OF MULTIPLATFORMISM

T. Verbitskaya and T. Korolyova

State University, Odessa, USSR

The problem of interaction between perceptive and articular aspects in foreign language learning is investigated. The possibility of correct comparison of sounds in mother tongue and foreign language in these aspects is demonstrated and the method to define distinct criteria of typological identity is suggested. The experimental results show the effectiveness of the method of interaction and perception of foreign language under the conditions of bilingualism.

Lately linguistics treats the problem of language system interaction under conditions of bilingualism and multilingualism quite often. One of the most important aspects in this respect is the way the languages are used. The problem of interaction between the speakers whose mother tongue is not the same and of special bilingualism which occurs under particular circumstances when a language is taught in educational institutions. These two forms differ from each other by their motivation. In the first case the principal aim is to exchange information while the forms of expressing ideas are more or less neglected. In the second occasion the pupils' attention is directed towards mastering the language. In this situation the language system is "pumped" to influence of the pupil's mother tongue (MT) on the foreign language (FL) studied. In case of natural bilingualism a mutual effect of contact between languages can be noticed. Both forms are to be closely examined by linguists and specialists of adjacent sciences.

Natural multilingualism is characteristic for the situation which has taken place in the Soviet Union. When Russian and national languages interact, their influence upon each other is noticeable: the Russian language spoken by the representatives of other nationalities acquires quite definite phonetic properties which are connected with the phonological and phonetic characteristics of either of national sound systems. Such interaction results in specific "national" variants of Russian language. This variant is typical not only for the representatives of national language but for Russians living in the Republics. It has been noticed that the peculiarities of sound realization are caused by the Russian language used in the Republic but not the national language of the Republic.

It is known that the artificial form of contact between speakers can be realized in the context of the fulfillment of several factors. The analysis of interference problems in linguistics suggests to let the potentially expected forms that can appear under the influence of phonological relations be done by means of finding out identity or difference in any aspect of the language, i.e. typological comparison of languages. Under modern conditions, when international oral communication is being intensified, one of the practical tasks of linguistics is to reduce foreign accent which is regarded as a phonetic phenomenon. Foreign accent causes much trouble in communicating with native speakers, the heaviest aggravation is on phonetic level.

In linguistics one of the leading views on the problem of interaction between perceptive and articulatory aspects in the question of double interference as a source of accent under the conditions of mixed bilingualism because phonology of hearing conditions the phonology of speaking. In a number of works perceptive aspect is looked upon as a theoretical one. Thus perceptive basis (PB) occupies a place alongside with the articulatory one (AB). PB is a unity of adjacent patterned phonetic units and the rules of comparison with them.

Distinction in various languages can be explained by the presence of direct connection between perceptive & acoustic aspects. The experimental data illustrate the presence of tight correlation between perception and articulation [1]. There is much interest in the results of the experiment as speech sounds perception which was taken to be a basis for the conclusion that perception system is characterized by the dependance of certain forms of speech perception on primary impression of one's mind. Trubetzyk's [4, 5, 9] underlined that the perception of foreign language sounds is phonologically conditioned. It is not mean that a foreign language speaker interprets any unfamiliar sound as a known one, i.e. there is no sequence into a sequence of native language phonemes. On the basis of the experimental data available we can state that capibility of differentiating a larger number of sounds than the amount of sounds in his MT, nevertheless, this capibility of a person is also as the correlation of phonological relations. When perceiving the sounds which are absent in the MT the examination of typological comparison of languages does not always place them as phonemes belonging to the MT rather sub différence is possible. It grounds on the properties of the processing of the sound signals, knowledge of one or several capibilities of the examined as well.

Everything mentioned above testifies to the necessity to specify the traditional view on phonological and perceptive abilities of a man. In this connection a new view on the important task is to find out the phonetic characteristics which are used in the set of perception. At the same time it would be wrong to forget that perception and (re)production are the two sides of joint activity.
the language studied in the
space of bi-lingual (poly-
langual) results from the
contact of phonetic system
in MT and FL. Such “hurt-
ful” points can be ticked out on
the basis of their interac-
tion in contact languages. Thus,
and articulatory phono-
mes in MT and FL on the
principles of their acoustic
or articulatory similarity; one
can find the following defini-
tion of “phonemes in close
group”, “phonemes of relatively
close group”, “phonemes of
proverbial close group”;
“phonemes coinciding in MT
and FL” and the like.
In the classifications
of this kind one can see the
criteria that underlie the
act of typology—yet it is
difficult to understand
what is meant as phonemes,
coinciding in MT and FL.
It is quite doubtful that
identical phonemes can
exist in heterogeneous lan-
guages. This conclusion
about coincidence of the
whole number of English
sounds /i, /u/ with the corre-
sponding Russian phonemes
is not convincing, because
two things are ignored:
a clearly distinguishable
option in the system
of Russian consonant-
palatal/hard sounds which
do not occur in English
(German) language and
pho-
not-mutation, at
the same time the
a poor correlation
of various palatal-hard
phonemes and the
same in Russian consonant
because, firstly, the
largest amount of consonant
phonemes is involved in
this opposition and, secondly,
palatal-hard consonants
influence greatly the adja-
cent vowels and cause the
whole range of peculiar-
ities in allophones of vowel
phonemes in speaking.

Now let it just to the
point to say that when carrying
out a typological research
not only the existence of
this or that vowel group to
be taken into consideration but
the place it occupies in the
system of the poly-langu-
under discussion.
Taking the above-mentioned
fact into account we can’t,
consider such phonemes as
German a, e, f, and Russ-
ian /a, /e, /i, and /y/
as identical ones, though
there is some definite arti-
cularly salient difference
between them. The diffi-
culty of mastering the
identical German sounds
is caused by the following peculiari-
ties of Russian
articulation basis: 1) in position
/ non-front vowel or voice-
less fricative there is no additional
raise of the back part of the
tongue; 2) after fricative
vowel there is no addi-
tional raise of the middle
of the tongue; 3) in position
voiced fricative there is no
regulation.

The latter one causes such
difficulty for reproduction
because of presence of two
different processes in the
system of Russian, Ukrainian
consonant, on the
one hand, and German conso-
nant, on the other hand:
regressive assimilation in
voicing and progressive
assimilation in


Compare (Forman) and (Angi)an-
ments. In voicing down a
mixture to come to the conclu-
sion that there are no dif-
ficulties in the process
of voicing-invoicing in the
German language is not so
strong as in Russian or
Ukrainian (5.12), that is
why the difference of
German voiceless /t/ and
voiced /l/ consonants
gives rise to noises even in
inertial position (see, for exam-
ple, the results of percep-
tive tests show that the
assimilation of voiceless
to voiced is mastered
more widely than
the back-


The results of tests carri-
out give grounds to admit
this statement to be
true since the sounds which
have no analogues in MT are
differentiated better than
the ones having correspondence
in MT. It can be explained by the
listener’s group influence on
the perceptual signal into the
sound of story and more
is in MT identical as to its
acoustic characteristics,
when practice with the
bilingual stimulus the
should be given to the
data given in the


It is demonstrated
german sounds /i, /u/ which
have relative analogues in
Russian and Ukrainian lan-
guages are worse mastered
in perceptive and articula-
tory aspects than sounds
/3, /f/ which have no
analogues. There is a clear-
yourly shown that the
sounds of the first group to
increase the amount of
error in sound perception
and articulation particula-
ly, when purposes
training in oral perception
and production of sounds in
the second group are much more
stable.

The use of the criterion
of comparability/innocen-
tility in articulatory acti-
tion at the whole sound
production allows to reveal
a clear-cut correlation be-
tween articulatory and
perceptive aspects of speech
under multilingual condi-
tions.

[ILLESKERK, A.M. (1967),
"Perception of Sounds Code",
Psychological Review, 35-68.
"Phonology at sound source
teaching foreign language pronuncia-
tion in speech. History of national
phonology Moscow, 566-71.
toetics of the French lan-
guage", Moscow.
basis of phonology", Moscow.
ology of language and
writing in Russian and
German. Leipzig:VEB Verlag
Enzyklo-

THE INFLUENCE OF SOCIAL FACTORS ON URBAN SPEECH

L.V. Ignat'kina

Leningrad State University, USSR

ABSTRACT

The influence of social factors (education and profession) on urban speech in twenty cities of Russia is discussed.

One of the important research areas in modern linguistics is the study of the standard variants of a national language and the factors with influence modification of a speech sound. This problem is closely connected with the study of the development of standard pronunciation, geographical variability, and social factors of speech phenomena. All those questions may be answered in the best way by the research of urban language. This paper is done in the line of macroso sociolinguistics, using F. Bell's definition and it is a part of the sociolinguistic research fulfilled in the USSR, Poland, Chechoslovakia, Germany and the USA by the linguists of different countries. The main attention is devoted to the factors of education and profession and their correlation with non-standard dialectal and low-standard language phenomena in the urban speech. The 3 following problems are being solved in the paper:

1. Fixing correlation between the regional speech features and educational level.
2. Finding out the influence of the "specialty" factor on persons speech.
3. Comparison of the speech features of representatives of different dialect zones (North-, Middle-, South-Russian, and the Ural and Siberia). The research provides additional material for the description of the socio-linguistical influence both on the standard and regional variants and helps to re-examine the functioning and the development of the orthographic norm. The analysis of the oral urban speech shows the factors of democratization of the Russian standard pronunciation, which is put to life mainly through urban speech in the processes of contacting with standard language and other forms of national language (local dialects, popular speech).

The towns and cities observed are situated within the zone of functioning of the Northern Russian dialects, Krasnodar, Kursk, Rostovon-Don, Ryazan, Simferopol, South-Russian dialects, Volgograd, Nizhny Novgorod (Gorky), Samara (Kubinsk), Pskov, Yaroslavl - Central Russian dialects, Nizhny Tagil, Novo sibirsk, Omsk, Tomsk, Sverdlovsk, Chelyabinsk. The speech is fixed in the Ural-Siberian group, Leningrad, as it is known, doesn't belong to the zone of functioning of any local dialects.

Analysing the speech of people living in these cities we have an opportunity to observe the effect of local dialects and popular speech on the standard language, as well as to find out the correlation between their frequency and the level of education (secondary in complete higher) and also the profession (philologist and engineer). The speech of those citizens, who were chosen from the natives of the city, from 18 to 60 years of age, which had higher or primary education or who were students. The speech of 20-30 people was recorded in each town. In Leningrad 150 people of social and other professions were recorded (21% (126 people) of the total quantity of the subjects had higher education, 14% (81 people) - secondary education, 65% (381 people) - were students; 232 philologists and 204 - of other professions.

The experimental text was phonetically representative, compiled of 3000 phonemes with regard to the most frequent combinations and positions. The texts were read by the subjects and tape-recorded, then analyzed mainly by ear. The results have shown that in the Leningrad speech there is significant difference between the phonetic units caused by the level of education. But the speech of the people with higher education is slightly closer to the ideal standard, than that of the people with secondary education.

We may speak about the more stable and more frequent character of the reproduction of the standards features only as about a tendency: that is the lack of occlusion during the pronunciation of the obstruents /s/, the lack of disimilation in the consonant cluster in the word /'l'ixko/ read as /'l'ikxo/ in the speech of secondary educated subjects. The frequency of mistake in each case gains 20%.

In the speech of other citizens there is a clear correlation between the presence of subnormal features of pronunciation and the level of education: the higher the frequency of the popular and dialectal elements is - the looser is the level of education. It's remarkable that in the speech of the South Russian towns citizens not only the popular features are stable, (the same as in the speech of other towns' citizens), but also the different dialectal features, for example the pronunciation of the fricative [z] instead of the normally consonant /s/ in the speech of the subjects from all the towns, except Southern, popular features are 2-3 times more frequent than dialectal ones. The simplification of the final consonant groups such as /'s'/, /'s'/, /pav exnas'/, /'zas'/ is widely spread everywhere. In all the cities, except Leningrad, the pronunciation

414

415
of students is to a larger extent more orphoeic than the speech of the subjects with secondary and even higher education. It seems to be explainable, by the fact that the students of regional high schools have a stronger desire to speak correctly. Hence, being the socially progressive group of population, the students of different profession were chosen as the subject of the further research.

The data on the typical deviations from norm are presented in Table. The percentage of philologists and subjects of other professions grouped according to the regions is the following: in the North-Russian cities the philologists comprise 9% from the total quantity of the students, students of other professions - 6%, in Uralo-Siberian cities: 11 and 23% correspondingly, in Middle-Russian cities - 12 and 16%, in South-Russian cities - 10 and 17%.

<table>
<thead>
<tr>
<th>Cities</th>
<th>Profession</th>
<th>Deviations of pronunciation</th>
<th>[e] [o]</th>
<th>/i/-/j/</th>
<th>/x/-/[x]</th>
<th>/k/-/x/</th>
<th>/o/-/o/</th>
<th>/d/-/[a]</th>
</tr>
</thead>
<tbody>
<tr>
<td>North-Russian</td>
<td>Philologists</td>
<td>7</td>
<td>6</td>
<td>22</td>
<td>58</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>Philologists</td>
<td>10</td>
<td>16</td>
<td>7</td>
<td>26</td>
<td>92</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>Others</td>
<td>12</td>
<td>3</td>
<td>4</td>
<td>12</td>
<td>25</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Uralo-Siberian</td>
<td>Philologists</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>17</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Siberian</td>
<td>Philologists</td>
<td>5</td>
<td>9</td>
<td>20</td>
<td>35</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>Others</td>
<td>9</td>
<td>0</td>
<td>13</td>
<td>43</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>South-Russian</td>
<td>Philologists</td>
<td>5</td>
<td>11</td>
<td>32</td>
<td>12</td>
<td>45</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>Others</td>
<td>9</td>
<td>10</td>
<td>23</td>
<td>26</td>
<td>52</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

On the basis of the given pronunciation are to a larger extent peculiar to the speech of non-philologists.

5. The percentage of the appearance of /æ/ in the unstressed position is relatively small. It is explained by the fact that the city is situated, and also of popular speech, which is not limited. Thereby the frequency of the dialectal features as a rule, is lower than popular ones, except the case with the South-Russian cities where fricative [j] is pronounced instead of occlusive [g] and [x] instead of [k] in the absolute final position.

2. The reproduction of vowels in all the cities is closer to standard than of the consonants.

3. Popular features, caused in general by casualness and passiveness of art reproduction, are more frequent in the speech of non-philologists. In North-Russian cities the students use popular elements more frequently than in other regions.

4. Dialectal features of the South-Russian dialectal features to the urban speech is also possible. The variability of pronunciation standard is supported by the dissemination of the popular features.

Summary. The speech of 888 people living in 20 cities of Russia is analyzed according to the weight of popular speech and dialectal factors. The cities comprise 4 groups with one and the same dialect in each group. Popular speech elements prevail in all the cities except the South-Russian ones. Sociolinguistic factors are also discussed: the level of education and the profession. It turned to be that the speech of the students in all the cities except Leningrad is closer to the standard, than that of the people with higher education. Philologists show a more correct speech than the students of other professions.
REGIONAL VOICE QUALITY VARIATION IN SWEDEN

Claes-Christian Elert and Britta Hammerberg

Department of Linguistics, Umeå, Sweden, and Department of Phoniatrics and Logopedics, Karolinska Institutet, Stockholm, Sweden

ABSTRACT

Recordings of c. 60 (incl. some bidialectal) speakers, representing important regional varieties of Swedish, have been subjected to spectrography. F0 (mean, range, distribution) and perturbation analysis, LTAS and listener panel evaluation are being planned. Preliminary findings are that Växjö (S. Sweden) voices are characterized by higher pitch, wider range and a overall high frequency creak. Gothenburg (W. Sweden) speakers use smaller jaw opening and exhibit phrase terminal creak.

1. GENERAL PRESENTATION

Voice quality is defined as the perceived overall characteristics of a speaker's speech. It depends on the morphology and size of the speech apparatus of the speaker and his articulatory habits. There are various ways of analyzing and describing voice quality. Laver [9] makes use of the concept of "articulatory setting", introduced by Honikman [6]. Voice quality is characterized by supralaryngeal settings, such as labial protrusion, pharyngolaryngeal, raised larynx etc., and phonatory settings, which are described partly in perceptual terms, such as falsetto, chest, harshness etc. Another approach to the analysis of voice quality is to study the relationships between the acoustic data and perceived voice characteristics, for instance by comparing rough, coarse, steady and nasal voice [5].

Voice quality has a wide range of linguistic and sociolinguistic functions. It can characterize a speaker's sex, age, personality, mood or relationship to a speaking partner. It distinguishes also groups of speakers, e.g. a language community as a whole ("Gesamtpräge" or "einselnen Sprachen"; [7, 246-251. In certain dialects [3] or sociolects [2], [10]. The regional variation of voice quality in Swedish and in most other languages has not been studied systematically. There are a few cursory references in the dialectology literature. A study of the sentence intonation in various parts of Sweden revealed that, on average, speakers from the north used a lower fundamental pitch than those from the south [8, 185]. A brief account of the pronunciation of voice quality features as high pitch, nasality, breathiness and creak in regional variants of Swedish is given by Elert (1) (with maps).

2. METHODS

The present paper reports the preliminary findings of an investigation of voice quality among speakers in selected areas of Sweden. So far recordings of text readings and spontaneous speech have been made by 60 speakers (men and women) from Gällivare-Malmberget, Götteborg, Linköping and Växjö. There are plans to obtain recordings of speakers in Stockholm and Umeå. The places have been chosen as representing important varieties of Swedish or interesting aspects of voice quality. It has been difficult to find subjects who are truly representative of the regional voice quality and to neutralize the effect of variation of individual voice properties among speakers of the same regional variety. We hence also compared recordings of dialect and standard Swedish by a few bidialectal speakers.

The recordings have been analyzed acoustically by various methods. Fundamental frequency distribution analysis (FEDA) yields, besides the distribution histogram, data, such as mode, mean and range of the fundamental pitch (F0) (in Hz and in cents). Perturbation measurements give values for small variations from period to period in the speech waveform. Spectrograms and oscillograms of part of the recordings have been studied. It is our plan to carry out long time average spectrum (LTAS) analysis. All these methods have been tested in the analysis of pathologic voices where they have yielded results which are highly correlated with perceptual categories of voice quality (see [5]). We have made a perceptual analysis of all recordings. Our plan is to supplement this analysis by submitting comparable portions of the material to an group of independent listeners for evaluation.

3. PRELIMINARY RESULTS

Average F0 is higher among the Växjö men than among comparable groups in Götteborg and Linköping. The pitch of Linköping men is not only lower but has also a smaller range. A general high frequency creak of most of the voices of Växjö subjects is easily perceived in an auditory analysis. Higher pitch and raising of the larynx was observed when a bidialectal speaker changed from a speech form close to Gothenburg Swedish to his native southern Småland dialect. Some irregularities in the waveform may be correlated with the properties perceived in Växjö voices (see Figs. 1 and 2). Another characteristic of the Växjö speakers is the overall velarization or uvelarization which is associated with the occurrence of uvular [R] or [r] (in medial and final position) a central or back vowel as allophones of the frequent phonemes /r/. Acoustic correlates of such features may be detected in a projected long time spectrum analysis.

The particular voice quality characteristics of the Linköping and Göteborg speakers are less clear. This applies both to the perceptual and acoustical analysis. The Göteborg speakers exhibit various forms of creak, esp. at the end of phrases. There is a preference among Göteborg male speakers to speak with smaller jaw opening.

4. REFERENCES


Figure 1. Waveform of [a(r)] and spectrum of [are] in the word *tätare* pronounced by the male Växjö speaker GP.

Figure 2. Waveform of [ar(e)] and spectrum of [are] in the word *tätare* pronounced by the male Göteborg speaker CO.
ETUDE SUR LA PERCEPTION DE L’"ACCENT" REGIONAL DU NORD ET DE L’EST DE LA FRANCE

Fernand Carton
Institut de Phonétique, 21 bd Albert 1er, 54000 Nancy
Robert Espresser
Institut de Phonétique, URA 261, 23 av. Robert Schuman, 11621 Aix-en-Provence
Jacqueline Vaissière
Institut de Phonétique, URA 1027, 19, rue des Bernardins, 75005 Paris

RESUME
Cette communication concerne des expériences de perception sur "l’accent" régional du Nord et de l’Est de la France. L’avant-dernière syllabe de phrases prononcées avec un "accent" régional a été permise avec l’avant-dernière syllabe de phrases correspondantes prononcées en français standardisé, et vice-versa. Les résultats confirment l’importance des événements acoustiques dans la perception de "l’accent" régional. Un second test visé à juger de l’importance de la courbe médiétique dans la perception de "l’accent". Les phrases prononcées en français standardisé et les phrases avec "accent" patoisant ont été synthétisées avec le fondamental (Fo) plat. Les résultats montrent que les variations de Fo ont une influence limitée sur le jugement des auditeurs.

INTRODUCTION
On dit couramment en France qu’un Tel à l’institution traitant des Lorrains et des Comtois, et qu’un autre à “l’accent” chantant de Marseille. Mais il est bien difficile d’établir objectivement ce qui paraît simple, et évident à l’oreille [5] et de caractériser les indices qui déclenchent chez l’auditeur la perception d’un "accent" régional. Les raisons de ces difficultés sont multiples. Tout d’abord, le degré d’"accent" perçu par les auditeurs est continu (d’où l’expression populaire de: "Il n’a pas d’accent", à un peu d’"accent", et "Il a un fort accent"). Ensuite, l’identification d’un "accent" peut être déclenchée par un événement précis dans le continuum vocal, par la réalisation acoustique partielle d’un seul phonème, par exemple la

diphongaison d’une voyelle ou la prononciation d’un /r/, ou les indices peuvent se répartir sur un domaine plus large de la phrase, voire sur la phrase entière. De plus, une combinaison d’indices non spécifiques peut devenir une marque d’une région particulière. Enfin, la perception des auditeurs varie en fonction de leur propre passé linguistique, qui n’est jamais uniforme.

La disponibilité d’éditeurs de signal conviviaux, les performances des nouvelles techniques pour modifier la fréquence du fondamental, la durée et d’intensité de voix naturelles permettent d’espérer apporter des contributions nouvelles aux études sur les accents régionaux. Les "accents" marqués peuvent être "débarrassés" par touches successives des indices qui constituent la marque. De la même façon, on peut se donner pour objectif d’établir toutes les règles de transformation (segmentales, prosodiques et autres) nécessaires pour "traduire" une phrase orale standardisée en phrase d’autre "accent" marquée régionalement. Ce type d’études n’est pas seulement utile pour compléter notre connaissance sur les accents régionaux et les productions verbales, mais pour mieux connaître "la norme" sous-jacente au jugement des auditeurs. En effet, la difficulté d’identifier les problèmes à résoudre pour améliorer la parole synthétique "standard" montre qu’on n’a pas encore fini de découvrir (malgré les progrès réalisés) tous les critères perceptifs utilisés par les auditeurs pour juger du naturel de la voix.

L’oreille est parfois très tolérante et parfois très exigeante et on se sent souvent démon pour expliquer les réactions des auditeurs.

I. RÉSULTATS D’EXPÉRIENCES PRÉLIMINAIRES
Des dizaines antérieures, acoustiques et perceptives, portant sur des phrases prononcées en milieu naturel du Nord, de l’Est et de l’Ouest, ont mis en évidence des caractéristiques (trois dernières syllabes) informatives et temporelles, indices possibles d’identification socio-geographique, le point culminant se situant sur l’avant-dernière syllabe du groupe. La rupture médiétique et sonore corrélée [1,2,3], la durée vocale (durée accente de l’avant-dernière "pénultième" syllabe calculée par rapport à la durée moyenne des syllabes accentuées [4,5]), et l’énergie et la durée consensuelles dans cet ordre ont montré qu’elles jouaient un rôle priori-mordial (voir aux [5,6,7]). Il nous a semblé intéressant de tester deux de ces points, le premier concernant l’importance de la pénultième, et le second l’apport des variations de Fo (en rapport avec la rupture médiétique). La permutation de la syllabe de la pénultième de phrases standardisées est-elle un indice nécessaire et/ou suffisant pour déclencher chez l’auditeur la perception d’un accent régional? (Test 1). La rupture médiétique et ses variations de Fo (suppression de toute rupture médiétique) n’eut-elle réellement à la perception d’un "accent" régional? (Test 2). Notre choix a été porté sur l’indice de phrases où "l’accent" régional est fortement marqué, afin de faciliter les tests de perception.

2. CORPS ET LOCUTEURS
Lors d’un enregistrement préliminaire, deux locuteurs, un conteur patoisant de l’Est de la France, de la région de Nancy (Locuteur L1), et un acteur patoisant du Nord de la France, de la région de Tourcoing (Locuteur L2), ont prononcé la même liste de phrases. L1 et L2 ont été choisis à cause de leur "accent" naturel et, aussi parce qu’ils pouvaient imiter le français standardisé. Les 22 phrases de la liste étaient des phrases très courtes du type “Ce n’est pas mon pitiou”. La pénultième de toutes les phrases comportent une syllabe ouverte correspondant à la consonne sourde /p/ et suivie de la consonne sourde /f/ et toutes les voyelles du Français possibles dans cette position sont représentées. Le choix d’un consonne sourde a été choisi afin de simplifier le problème de l’extraction de la voyelle pour l’épreuve de permutation. Nous aurions voulu fixer complétement un plus large contexte phonétique, afin que les phrases diffèrent seulement par l’identité de la voyelle de la pénultième, mais il n’a pas été possible de trouver des phrases neutres satisfaisant ces conditions. Lors d’une première écoute, on avait également noté que les phrases à "accent" avaient tendance à avoir une forte connotation expressive, alors que les phrases en français standardisé étaient prononcées de façon relativement neutre. Un autre enregistrement subdivisé par d’autres locuteurs passant (dont un phonéticien) ont confirmé cette tendance générale. Ce problème de degrés différents d’expressivité n’a pu être résolu.

Les phrases ont été répétées 4 fois par l1 et L2, deux fois en laissant, le soir au locuteur de la marque régionale, et deux fois en français standardisé. Parmi les 176 phrases (22 phrases * 2 styles * 2 répétitions * (L1 + L2), 10 phrases par locuteur où l’"accent" régional était le plus apparent ont été sélectionnées, ainsi que les phrases sans "accent" correspondantes (20 phrases en tout). Ce sont ces 20 phrases qui ont servi aux tests.

Les auditeurs ont été séparés en deux groupes selon leur langue maternelle: 20 auditeurs français, essentiellement de la région parisiennne et 20 auditeurs parlant français et vivant actuellement à Paris, tous étudiants à l’Institut de Phonétique de Paris III (niveau licence).

TEST I: LA PENULTIÈME
Pour tester de l’importance de la pénultième, 4 types de phrases ont été présentés aux auditeurs: les 10 phrases standardisées (“Ce n’est pas mon pitiou”: S1), les 10 phrases correspondantes à "accent" patoisant (“Ce n’est pas mon pitiou”: PP), les phrases standardisées où la pénultième a été permise, après normalisation pour rendre l’intensité de la syllabe égale à celle de la syllabe à remplacer, avec la pénultième des phrases correspondantes à "accent" patoisant (“Ce n’est pas mon pitiou”: SP), les phrases régionales où l’émission de la voyelleée sourde /p/ et suivie de la consonne sourde /f/ et toutes les voyelles du Français possibles dans cette position sont représentées. Le choix d’un consonne sourde a été choisi afin de simplifier le problème de l’extraction de la voyelle pour l’épreuve de permutation.

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1/1</td>
<td>2/1</td>
<td>3/1</td>
<td>4/1</td>
<td>5/1</td>
</tr>
<tr>
<td>L1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>165</td>
<td>153</td>
<td>196</td>
<td>150</td>
<td>199</td>
</tr>
<tr>
<td>PP</td>
<td>271</td>
<td>254</td>
<td>276</td>
<td>296</td>
<td>371</td>
</tr>
<tr>
<td>%</td>
<td>30%</td>
<td>35%</td>
<td>40%</td>
<td>45%</td>
<td>52%</td>
</tr>
<tr>
<td>L2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>170</td>
<td>151</td>
<td>156</td>
<td>223</td>
<td>182</td>
</tr>
<tr>
<td>PP</td>
<td>289</td>
<td>307</td>
<td>293</td>
<td>274</td>
<td>284</td>
</tr>
<tr>
<td>%</td>
<td>32%</td>
<td>62%</td>
<td>63%</td>
<td>21%</td>
<td>53%</td>
</tr>
</tbody>
</table>

Tableau 1: Durée de la pénultième (en ms) dans les phrases standardisées (SS) et les phrases patoisantes (PP) prononcées par le locuteur de Nancy (L1) et celui de Toulouse (L2) et pourcentage d’augmentation de la durée.

Les courbes de Fo des phrases patoisantes indiquent des déviations importantes par rapport aux courbes morphées couramment attestées en français standardisé, où Fo descend de façon régulière sur les dernières syllabes des phrases, à partir de la fin de la dernière syllabe de l’avant-dernier mot. On a noté deux formes typiques. La première consiste en un Fo bas sur la pénultième, suivi d’un rehaussement de Fo sur la dernière syllabe. La deuxième consiste en un ton montant sur la pénultième, suivi d’une valeur haute sur la dernière syllabe. Les deux contours expriment des degrés différents d’expressivité. Dans les deux cas, la dernière syllabe a une valeur de Fo plus élevée que la pénultième, ce qui est en contradiction avec le schéma final descendant des phrases standardisées.

RESULTATS

Les auditeurs ont eu à juger (jugement force) si chaque phrase entendue possède “pas d’accent” (note 0), “peu d’accent” (note 1) ou “un fort accent” (note 2). La tâche a été considérée comme facile par les auditeurs français et étrangers. Le tableau 2 ci-dessous indique les résultats.

Comparons les tableaux 2a et 2b. On remarquera que 90% des phrases standardisées sont perçues comme n’étant pas ou peu marquées par les sujets français, mais seulement 65% par les sujets étrangers. Dans 15 % des cas, les étrangers perçoivent comme fortement marquées des phrases standardisées (contre 1% des cas pour les français). Par contre, les sujets étrangers perçoivent comme sans accent 11% des phrases patoisantes (contre 2% des sujets français). La plupart de ces écarts sont dus au manque de méthode mis à leur disposition pour améliorer leur propre “accent” étranger, et des tests de ce genre semblent confirmer leur perception faible de l’accent.

Quel est l’effet de la permutation de la pénultième? La majorité (63%) des phrases standardisées où la pénultième a été remplacée par une syllabe extrait des phrases “patoisantes” sont perçues par les sujets français comme étant fortement marquées. Cela confirme le rôle important joué par la pénultième. L’étude cas par cas des phrases montrent que c’est l’introduction de syllabes nasales patoisantes (relativement peu nasalisées avec l’accent régional), dans des phrases marquées régionalement et de la voyelle postérieurisée qui est le plus efficace. Des tests en cours permettront de quantifier l’apport de chaque “écrit” de prononciation par rapport à la norme et d’expliquer les cas où la permutation s’est révélée inefficace.

Tableau 2: Résultats du Test 1 sur les auditeurs français (2a) et étrangers (2b). Sp représente les phrases standardisées où le pénultième a été permis avec la pénultième de la phrase à accorder correspondante. 0, 1 et 2, ici, notent les phrases qui ont été perçues “sans accent”, “avec un peu d’accent”, et “beaucoup d’accent”, respectivement.

Le rôle n’est cependant pas symétrique: il ne suffit de remplacer la pénultième d’une phrase patoisante et de la remplacer par une syllabe standardisée pour que la phrase soit perçue comme standardisée. En d’autres termes, il est plus difficile de débarrasser une phrase de son accent régional que de transformer une phrase naturelle en une phrase marquée. Dans la majorité des cas, la phrase peut être perçue comme étant plus que 34% ou fortement marquée (27%) et la permutation n’est efficace que dans 39% des cas. L’efficacité du changement varie en fonction des phrases restantes.

CONCLUSION

Le premier test a confirmé à la fois le rôle important de la pénultième dans la perception d’un marque d’une région particulière, et l’incidence d’autres facteurs. Le second test a montré que l’absence de rupture médiocque dans la clausule finale n’est pas une condition suffisante pour la perception d’une phrase standardisée. Ce dernier résultat nuance l’affirmation selon laquelle “l’intonation des français régionaux reste souvent la seule indication d’accent par rapport au français standardisé” (16) (Pg 7). Ces deux tests suggèrent l’efficacité d’une approche par transformations successives et contrôlée de voix naturelle. L’analyse acoustique et perceptive d’un corpus, ainsi que le fait de pouvoir revoir le corpus, peut apporter une réponse définitive au problème de la combinaison des indices non spécifiques qui deviennent une marque. Cette hypothèse doit être confirmée par l’analyse de nouveaux moyens de tests (méthode PSOLA par exemple, développée au CNET et utilisée ici, considérablement supérieure aux méthodes plus anciennes, faites à partir de LPC) de nouvelles méthodologies pour l’étude des marques régionales, incluant également des transformations spectrales, deviennent possibles. La technique devance notre savoir: saurons-nous en tirer pleinement profit?

REFERENCES

THE EFFECT OF ADDRESSEE FAMILIARITY ON WORD DURATION

Jan McAllister, Cathy Sotillo and Ellen Gurman Bard

Human Communication Research Centre and Department of Linguistics, University of Edinburgh

ABSTRACT

This paper describes an experiment which was designed to test the hypothesis that speakers alter the forms of words in response to the degree of familiarity of their interlocutor; specifically, that words addressed to a hearer whom the speaker knew well are shorter than the same words addressed to a hearer whom the speaker had not previously met. Six of the eight speakers examined exhibited the predicted effect in both read and spontaneous speech modes.

1. INTRODUCTION

Many factors affect the durations of spoken words. While some of these relate to the word's position in the immediate context of the utterance in which it occurs (for example, its proximity to syntactic boundaries or pauses [1]), others relate to the word's position with its wider linguistic and paralinguistic context; in particular with the extent to which speaker and hearer share knowledge and assumptions: for example, a word's duration is inversely related to its predictability [2,3]; words are longer when they are initially introduced into a discourse than when they are produced by the same speaker reading back a transcript of the same discourse [4]. The experiment described in this paper was designed to investigate the effect of a further variable in this latter group: the degree of familiarity between two interlocutors engaged in a cooperative task.

The starting point of this study was the hypothesis that word durations would be shorter when the two interlocutors knew each other well than when the task involved two speakers who had never previously met. It seems likely that familiar speakers will respond to their hearers' ability to use knowledge about what they say and how they say it, and shorten words, in much the same way as they might exploit the redundancy in utterances like A stitch in time saves nine to shorten the final word [2].

It has indeed frequently been claimed that speakers alter their speech and language in response to their degree of familiarity with the hearer (e.g. 7). Indirect experimental support for the hypothesis comes from more than one source. One type of evidence is found in the literature on the processing of spontaneous speech (see, for example, 5, 6, 9, 10). In such studies, the spontaneous speech samples have generally been elicited by having the subject(s) converse with the experimenter or some other person whom they have never previously met. However, the pairs of speakers who produced the spontaneous speech in McAllister's study [8] were close friends (and thus highly familiar) whereas the follow-up experiment, common with other researchers who have studied intelligibility in spontaneous speech, McAllister found that intelligibility was mediated by word duration; however, the level of intelligibility of content words in her spontaneous speech samples was marked lower than that in other studies of spontaneous speech.

McAllister suggested that the degree of familiarity of the interlocutors in her materials may have affected the duration, and thereby rendered the intelligibility, of the words she examined.

Further indirect evidence for the influence of addressee familiarity on the forms of spoken words comes from the experimental literature on motherese, the specialised register addressed to children. Shockey and Bond [11] found that phonological rules such as palatalisation open vowels in motor's speech to their children than in their speech to an adult viewer. In their experiment, addressed age was confounded with addressee familiarity: the mothers who took part in the study spoke to their own children, and to another adult whom they presumably knew less well. This suggestion is in keeping with Shockey and Bond's own proposal that the effect they observed was attributable to the mothers' wish to set a tone of intimacy in their dialogues with their own children. Although their proposal was supported by a significant variable studied by Shockey and Bond was phonological rule application rather than word duration, it is implausible that the two variables might be subject to similar influences, and indeed a further study of motherese [12], in which addressed age and addressee familiarity were similarly confounded, revealed that words addressed to children were shorter (as well as less intelligible) than those addressed to adults.

2. METHOD

The spontaneous speech samples which were used in the current study were collected in the so-called map task [13], which involves pairs of speakers, each of whom has a map. One speaker, the follower, is asked to describe a route marked on his or her map, while the other, the instruction giver, has no route. The speakers are told that their goal is to represent the instruction giver's route on the instruction giver's map. Neither speaker can see the other's map, and if the route described in the task described in this paper, the speakers were prevented from seeing each other by the presence of a screen. A number of pairs did not identify each other in every respect 1 and speakers are told this explicitly at the beginning of their first session. It is, however, the words to the speakers to distinguish between the two maps. They are encouraged to ask as many questions as necessary in order to achieve this end. The task has been used extensively to study speakers' discourse strategies and is considered by experimenters and subjects able to elicit highly natural spontaneous speech.

The eight subjects who volunteered to take part in the experiment were grouped into two 'quadruples'. Each quadruple contained two pairs of speakers. The members of a pair knew each other well but had never before met the member of the other pair in their quadruple. Each subject participated in map conversations: once as Instruction Giver with the other member of their pair, once as Instruction Follower with the other member of their pair, once as Instruction Giver with a member of the other pair in their quadruple, and once as Instruction Follower with a member of the other pair in the quadruple. Each subject thus participated in two sessions in the familiar condition (in which they knew their partner well) and in two sessions in the unfamiliar condition (in which they were partnered with a subject whom they had never met prior to the experiment).

Each of the sixteen spontaneous conversations which resulted from these pairings was orthographically transcribed by one experimenter and the transcription checked by another. The eight subjects were then asked to return to the recording studio and 'act out' their original conversations by reading from the transcript. They were partnered in each conversation by the same person with whom they had originally taken part in the experimental session. These recordings gave rise to a set of read materials.

From the transcripts, twenty different word types were selected for each speaker. The words which were selected were all content words, and each word had been uttered by the speaker in question when addressing both the familiar and the unfamiliar addressee. As far as possible the items were chosen from the transcripts in which the subject was acting as Instruction Giver.

The location of the first occurrence of each of the four pairs (Spontaneous / Familiar; Spontaneous / Unfamiliar; Read / Familiar; Read / Unfamiliar) were noted and the durations measured using the ILS display, and a conventional acoustic landmarks to identify word onsets and offsets [1]. The results from the session were then based on the analysis of 640 word tokens: 8 speakers X 20 word tokens X 2 addresses (familiar, unfamiliar) X 2 versions (read, spontaneous).
3. RESULTS

Table 1 shows the mean duration of the words in the four conditions, for all eight speakers.

A three-way analysis of variance (Version X Addressee X Speaker) was conducted. Not surprisingly, differences between speakers were highly significant \( F(7,152) = 3.21, p = .0034 \), partly because of differences in the speech habits of particular speakers and partly because no attempt was made to match word types across speakers, resulting in a different number of one, two and three syllable words in each sub-sample. Similarly, a version effect was observed which was similar to that previously reported in the literature [3]; spontaneous tokens were longer overall than read tokens \( F(1,152) = 2.89, p = .0912 \), but it interacted with the Speaker variable \( F(7,152) = 2.80, p = .0091 \); further analysis by Scheffé test revealed that all but two speakers (1 and 2) exhibited the predicted Addressee effect for both read and spontaneous speech: that is, words were shorter when addressed to a familiar addressee than an unfamiliar addressee. In a subsequent analysis of variance of the durations of word tokens spoken by these six speakers, Addressee proved significant as a main effect \( p = .0033 \), and did not interact with either of the other variables.²

Table 1: durations of words (msec)

<table>
<thead>
<tr>
<th>Spkr</th>
<th>Fam Spont</th>
<th>Fam Read</th>
<th>Unfam Spont</th>
<th>Unfam Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>393</td>
<td>362</td>
<td>383</td>
<td>411</td>
</tr>
<tr>
<td>2</td>
<td>343</td>
<td>281</td>
<td>347</td>
<td>328</td>
</tr>
<tr>
<td>3</td>
<td>383</td>
<td>347</td>
<td>383</td>
<td>391</td>
</tr>
<tr>
<td>4</td>
<td>347</td>
<td>355</td>
<td>383</td>
<td>347</td>
</tr>
<tr>
<td>5</td>
<td>370</td>
<td>347</td>
<td>370</td>
<td>355</td>
</tr>
<tr>
<td>6</td>
<td>347</td>
<td>347</td>
<td>383</td>
<td>391</td>
</tr>
<tr>
<td>7</td>
<td>347</td>
<td>355</td>
<td>347</td>
<td>391</td>
</tr>
<tr>
<td>Mean</td>
<td>388</td>
<td>347</td>
<td>383</td>
<td>391</td>
</tr>
</tbody>
</table>

Address was not significant as a main effect \( F(1,152) = 2.89, p = .0912 \), but it interacted with the Speaker variable \( F(7,152) = 2.80, p = .0091 \); further analysis by Scheffé test revealed that all but two speakers (1 and 2) exhibited the predicted Addressee effect for both read and spontaneous speech: that is, words were shorter when addressed to a familiar addressee than an unfamiliar addressee. In a subsequent analysis of variance of the durations of word tokens spoken by these six speakers, Addressee proved significant as a main effect \( p = .0033 \), and did not interact with either of the other variables.²

4. CONCLUSION

The experiment described here offers some support for the hypothesis that speakers shorten words when conversing with people whom they know well. The majority of the speakers here exhibited the predicted effect. Further work is now in progress to examine a number of related issues. First, more data needs to be examined to see how generalisable these preliminary results are to a larger number of speakers. Second, a wide variety of factors is known to affect word duration, but given the nature of the elicitation task it was impossible to control for all of these. Pause location, speech rate and syntactic structure are among the variables we plan to examine; however, analyses we have already conducted show that the Addressee Familiarity effect remains even when word frequency and word length in syllables are taken into account. Finally, we wish to determine whether speakers alter other aspects of the form of spoken words in response to addressee familiarity; research is in progress to examine the effect of the variable on speakers' application of connected speech rules such as stop deletion (see [14]).

The support of the Economic and Social Research Council UK (ESRC) is gratefully acknowledged. The work was part of the research program of the ESRC funded Human Communication Research Centre (HCRC).

NOTES

(1) The design of the maps being used in a large-scale study of Scottish English is described in [14].

(2) It is interesting to note that the two speakers who failed to exhibit the Addressee effect were those whose performance differed from that of the other speakers in other respects; in particular, their conversations were over twice as long as those of other participants in this and other studies using the same task. It may be that their unusual attention to detail in the task led them to adopt inappropriate linguistic behaviours.

(3) See also Bard and Anderson [13] for a detailed analysis of this effect.

REFERENCES


ABSTRACT

The report deals with the description of orthoepic problems of Modern Russian Literary language and contains the results of experimental phonetic research for held on all lexical basis of Russian language. The work is fulfilled with the purpose of forming the Phonetic base of Russian improving other number of applied systems: automatic recognition and synthesis of speech, correct pronunciation training, and automatic transcription.

One of the most prominent trends in the development of Soviet linguistics recently is the creation of Computer data base of Russian language as a complete data base of phonetic system and functioning of Modern Russian Literary language. Phonetic part of the Computer data base suggests attaining and classifying knowledge of sound side of language taking into consideration all existing pronunciation variants. Prior to creating such phonetic data base number of complicated theoretical and practical problems must be solved. On the other hand existing phonetic data base will greatly enlarge the possibilities of applied use of phonetic data. Thus, the question of relations between norm and non-norm (is non-norm always a mistake and must dictors always have ideal pronunciation?). Problem of unique or multivariant orphoepic norm in different types of speech activity as well as the question of position of those phonetic systems which are realized in different types of speech (on different lexical material) and have their own laws of construction and functioning (many systems or one system with many subsystems?). On the other hand, creation of Phonetic data base of Modern Russian Literary language allows to improve such applied systems as automatic recognition of speech, synthesis and automatic transcription of Russian speech. Phonetic disciplines teaching - theoretical phonetics, Russian pronunciation and practical transcription - study of phonetic similarities of spontaneous speech and results of different interference processes, both between languages (Russian speech of Non-Russians) and inside one language.

For all mentioned above it is very important to find out existing pronunciation variants for all totality of Russian lexica, especially for peripheral parts of lexical system (borrowings, abbreviations, complex words and so on). Up to now only studies were held on the limited material, the task of receiving recommendations for each word was not put on. Now there is possibility to store the whole dictionary in computer memory and to treat them automatically.

Due to all these reasons a new series of orthoepic studies in which students of the philological department take part has been started in Leningrad state University laboratory of experimental phonetics named after L.V. Shcherba. All studies are experimental including methods of auditory, instrumental and psycholinguistic analysis.

Material in all cases is maximally complete - different Russian dictionari- es of new and foreign words, abbreviations and special lexica, frequency and derivative lexicon. In all cases the auditory material recorded by dictors - philologists who produced speech was tested and approved by special test. Words with orthoepic difficulties were put into phrases in identical syntactic positions. Auditors were students and researchers of philological department. Analysis was made by micro-computer of DKK-type (regeneration of auditory material, duration measurements, phonetic similarity measurement). Results in all cases are concrete recommendations in pronunciation and transcription as well as relations between found orthoepic variants. Some of these results are given below.

Among the words with complex consonant combinations those which contain combinations СТ, ЗДН, СТК, НСК (кошляны, бешем389483, берег) were studied. Complete lists of such words were selected from the "Russian Deriva- tional Dictionary" by D.Worth, A.Kozak and J.Johnson (New-York,1970). Further, R.DD, those for which existing orthoepic recommendations (N.A. Avan- ciev, L.A. Verbitskaya, modern orthoepic dictionari- es) were not enough or didn't exist at all, were included in experimental material.

Experiments showed that pronunciation of words with СТ depends on the route: in the words with routes -КСТ-, -ТСК, -ТК- (кошляны /stl/), (КСТ/ /stl/), (КСТ/ /stl/) all consonant variants are preserved in pronunciation; in other situations dieresa is observed - with /zn/ and /stl/ or with /zn/ and /stl/ or /zn/. Basing on the route it is easy to formalize the pronunciation rules of such words.

For words with 3ЕН combinations among two pronunciation variants - with дверца /zn/ and /zn/ and /zn/ this the first is clearly prevailing (from 85% to 97% realizations for different words).

Study of words with СТК, НСК and ВДК combinations showed three pronunciation variants: with дверца /zn/ and /zn/ assimilation in the place of origin /zn/ and /zn/ and without die-
The prevailing of first variant is rather considerable in all cases: from 75% (in word ноМСКпремь) to 98% of all realizations. In all other variants only full pronunciations of word ноМСКпремь (20.3%) must be taken into consideration without argument.

Words with АРО, АРЕ and ОМЕ also difficult for Russian pronunciation turned out to be borrowed and badly mastered by Russian native speakers. For these words three pronunciation variants were found: with strong /3/, with /I/, and completely without /3/. The last variant turned out to be relevant for words with АРО: 16% before the stressed /о/ - РАЮН, МАЙОНИЯ: 45% in unstressed combination - МАЙОНЕЗ, МАЙОРАТ. Two other variants must be taken into consideration in pronunciation teaching, transcription and other applied aspects.

Among words with untypical for Russian language vowel combinations a group of words with ЕО in the route was studied. All the words are borrowed and are of terminologic character. The pronunciation difficulty of such words is defined by two factors: first only 7% of such words have stress on the second component of the combination, in 93% it is totally unstressed and stands in 1 to 6 prestressed position in the word; second only 26.4% of words are known to Russian native speakers and are used by them in speech. Other 37.4% are known but rarely used, and 30.8% are unknown and totally unused. During the studies it was found out that for some words (архЕОЛОГИЧЕСКИЙ, ГЕОТРИЧЕСКИЙ and so on) along with two-component realization (auditors fixed /10/, more seldom /е/0/) the realization of second component as one vowel must be taken into consideration. In the latter case in first prestressed position the second component of combination - /8/, more seldom /о/ is recognized as a rule: in the second and further prestressed positions - first component /е/, more seldom /8/. The realization of stressed combination ЕО also turned out to be monovocal - in words МЕТОР, ГЕОФИЗАТУР, ГЕОГРАФ.

The validity of received results was in all cases was checked during the control experiment in recognition of studied combinations realizations and realization of specially selected Russian words with identical phonetic structure: сёзобный - звёздный, кунгурский - кунсткамерский, архЕОЛОГИЧЕСКИЙ, НАУЧНОСТИЙ, МЕТЕОРИЧЕСКИЙ, МЕТРОГРАФ.

The newest borrowings into Russian language among which 10% of words with possible violation of Russian pronunciation norm were found are especially interesting for the studied problem. All in all 0% borrowings taken from different dictionaries of new words were studied. 56.5% of these words are on the first stage of mastering: tested philologists never met these words and didn't know their meaning. Only 9.3% of words are actively used by native speakers (колянхэ, кёнс, аэробика and so on). 32.6% of word from the list may have a hard consonant before orthographic Е - /брэйк, вебс/на/ и 22% - unstressed /о/ - консоме, банкн/но/, 10% - long consonants outside a morpheme connection /каптэ́йн/.
Le division de la chaîne paroles en sons directs relève des procédés phonologiques [8,9,10]. Il n'en suit que dans les langues de différents types en trouve des sons dont le statut linéaire n'est pas le même et qui répondent aux caractéristiques phonologiques partielles [5]. Depuis la classification universelle des sons du langage est possible grâce à l'unité des mécanismes de production et de perception de la parole. Pour stabiliser une telle classification on doit tenir compte des possibilités articulatoires de l'homme, ce qui permet de prévoir des cas non seulement pour les sons déjà connus, mais aussi pour ceux qu'on pourrait trouver dans les langues non étudiées [1,10]. Ce principe fut appliqué il y a cinq ans à l'établissement du tableau des voyelles par H. Sweet [7] et développé par S. Jones [2] et O. Jones [6] pour des voyelles cardinales de D. Jones, en forme de trapèze, qui est proposé aujourd'hui, presque sans modifications, par l'Association Phonétique Internationale [6] en réfère plus au moins approximativement les caractéristiques phonologiques. La forme du tableau et les points de repère (choisis d'une manière conventionnelle et qui ne coïncident avec aucune voyelle réelle [5], bien qu'étant en corrélation avec certains phonèmes de plusieurs langues) permettent de trouver la place pour n'importe quel son vocalique et, par conséquent, tenir compte des oppositions phonologiques éventuelles, qu'on pourrait en perfectionnant ce tableau, on se borne à préciser la localisation d'un son et à choisir un symbole répondant mieux aux besoins pratiques.

Quant aux consonnes (il n'est pas dans la suite que des consonnes palatalisées), la situation est différente. Contrairement aux voyelles, qui sont placées le long des axes représentant des caractéristiques continues, le classement des consonnes se fait sur la base des caractéristiques discrètes, dont les gradations doivent être établies à l'avance. Conformément à la tradition, on a 3 gradations pour le mode d'obstruction (ou 4, si on considère les affriquées comme un type à part); le nombre des points d'articulation varie dans différentes systèmes de classification (Ginder [10] en compte 13 et l'API [6] en compte 16 et 1/2). On distingue aussi 3 types de consonnes selon le degré de sonorité. Ces trois caractéristiques servent à décrire l'articulation consonantique de base (qu'il faut distinguer des articulations onomatopéiques ou effectuées par les organes qui ne participent pas à l'articulation de base, p.ex. labialisation des consonnes non labiales), elles fonctionnent en même temps comme traits distinctifs de phonèmes. Donc, les limites entre les 3 types de consonnes sont établies sur la base des oppositions phonologiques, comme le prouve P. Ladefoged [5]. Cependant ce principe n'est pas réalisé d'une manière conséquente et en plus est discutable du point de vue phonétique. Les oppositions phonologiques ne sont guère universelles. On pourrait trouver dans une langue non étudiée une opposition phonétique non prévue par la classification (p.ex., deux types de médialinguales qui diffèrent grâce à l'organe passif, ou les postlinguales). D'autre part, il y a des types consonantiques qui ne forment pas d'opposition, p.ex. [m] et [ŋ], et dont l'articulation de base est différente. Le tableau de consonnes présenté donc un compromis entre une classification selon les oppositions phonologiques et celle basée sur les caractéristiques articulatoires.

Si on veut suivre le principe d'universalité et prévoir la possibilité de classer toutes sortes de sons dont l'articulation de base est différente, on peut compléter et modifier le tableau de consonnes de l'API. Le tableau modifié est présenté à la page suivante. On a exclu certaines consonnes implosives: les consonnes éjectives ou éjectées ou bien comme ayant deux points d'articulation (ou même titre que [k] par exemple, puique [?] est classé comme les consonnes occlusives) ou bien comme glottalisées, c'est-à-dire caractérisées par une articulation secondaire. Pour les principes de classement on a introduit dans le tableau les affriquées (pour des raisons techniques quelques symboles sont omits); on a indiqué la possibilité de réaliser toutes les fricatives comme sonnantes. Les oppositions palatalisées selon l'organe articulatoire actif; dans certains cas l'organe passif doit aussi être pris en considération (p.ex., pour les labiales et les postlinguales).
Classification des consonnes (d'après l'articulation de base)

<table>
<thead>
<tr>
<th>Mode d'obstruction</th>
<th>Labiales</th>
<th>Linguo-labiales</th>
<th>Dental-</th>
<th>Alvéolaires</th>
<th>Palatinas</th>
<th>Médio-palatinas</th>
<th>Post-alvéolaires</th>
<th>Antépalatotherales</th>
<th>Pharyngales</th>
<th>Scotiales</th>
</tr>
</thead>
<tbody>
<tr>
<td>occlusives pures</td>
<td>p b</td>
<td>p b t d t j t d c k q g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>occlusives bruits</td>
<td>b v</td>
<td>t z d z s z s z s z</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fricatives latérales</td>
<td>m m m n n n n n N</td>
<td>m m m n n n n n n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fricatives médianes</td>
<td>w v</td>
<td>z z z z z j x y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fricatives latérales</td>
<td>w v</td>
<td>z z z z j y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>roulées médianes</td>
<td>f r</td>
<td>r r r r r r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bruits et consonnes latérales</td>
<td>f r</td>
<td>r r r r r r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Suivant L.Ščerba et L.Žinder[10] on a indiqué l'existence d'articulations uvulaires faucales: le son se forme au moment où la luette se détache brusquement de la paroi pharyngale; russe днe "fond", anglais sudden etc. On ne connaît pas d'opposition phonologique de ces consonnes, mais théoriquement cela n'est pas impossible. Au lieu d'un seul groupe de pharyngales de l'APhI, le tableau en contient deux: supérieures, formées au niveau de la racine de la langue, et inférieures, articulées avec la participation de l'épiglotte; ces deux types de consonnes existent comme phonèmes indépendants dans certaines langues du Caucase.

Le tableau présente une classification plus détaillée de fricatives médianes, qui sont divisées en trois groupes selon la forme de constriction: consonnes à fente ronde (p.ex. [s],[w]), consonnes à fente plate (p.ex. [θ],[ʒ]), consonnes à fente allongée vers le palais mou, à cause d'une plus grande élévation de la langue - les chuintantes. Les trois articulations ne sont différenciées que dans certains groupes de prélinguaires.

Les vibrantes sont divi-sées en médianes et latérales. Les médianes sont articulées avec la pointe de la langue ou la luette. Les latérales sont formées grace aux vibrations des bords de la langue avec la pointe pressée contre les alvéoles. L'opposition phonologique entre ces deux types de consonnes est peu probable; mais leur pré-

sence dans le tableau est justifiée par le principe d'universalité qui oblige à tenir compte de toutes les articulations possibles.

REFERENCES
[8] Ščerba, L.V. (1912), "Rousskie glasnye v katkestvennom i kolitchestvennom otbochenâi, St-Pétersbourg.
GEMINATION PHONETIQUE EN FRONTIERE DE MOTS
A. Marchal & A.S. Del Negro
URA 261 CNRS, Parole et Langage, 29 Av. R. Schuman
13621 Aix en Provence

ABSTRACT
This paper is concerned with the production of identical consonants at word-boundaries. The question which arises is whether to know if these consonantal groups result in one or two articulatory gestures. We investigated in an EPG and acoustic study 3600 cases of stops (1200 single consonants and 2400 pseudo-geminates; 10 speakers, 10 repetitions of 36 natural sentences). Our data clearly indicate that these groups are produced as a single long consonant and that there is no evidence of recategorization during the closure phase.

1.INTRODUCTION

En Francais, les géminées apparaissent au 10ème siècle se sont simplifiées en consonnes simples [2]. Bien que la graphie de consonnes doubles ait subsisté où a été empruntée (mots savants, reconstructions étymologiques), les géminées n’assurent plus de fonction distincte à l’intérieur d’un mot.

D’un point de vue phonétique, un problème intéressant est posé par le cas de la rencontre de deux consonnes identiques à la frontière de mots. Sont-elles réalisées comme deux consonnes distinctes ou comme le prolongement d’un même geste articulatorièr ? Le cas échéant ; comment les distinguer d’une consonne longue?

Pour résumer, les questions que l’on peut se poser sont les suivantes.

- Au niveau acoustique, les groupes de consonnes identiques en frontières de mots se comportent-ils comme une consonne simple mais longue, ou bien comme deux consonnes distinctes?

Pour trois des locuteurs, l’acquisition numérique simultanée des données acoustiques et électropalatographiques est réalisée à l’aide de la station PHYSIOLOGIA ACCOR [5].

3.MEASURES

L’appui linguo-palatal exprimé en nombre de contacts par zone articulatoire (total, antérieur, postérieur) est mesuré pour les trames suivantes :

- 1ère trame d’occlusion complète (C1, C2 ou X);
- trame de maximum de contacts (C1M, C2M ou XM);
- trame précédant tout relâchement (R1, R2 ou RX).

4.RESULTATS
4.1 La réarticulation
Les principales observations tire de l’examen des tracés palato-graphiques et de l’analyse détaillée du signal acoustique font apparaître:

4.1.1 Pour les géminées voisées
- Une tenue stable
- L’absence d’un seul mouvement articulatoire
- La persistance du voûtement pendant toute la consonne

4.1.2 Pour les géminées sourdes ou assourdies
Dans le cas des consonnes sourdes ou assourdies, il faut distinguer ce qui se produit pour les palatales des phonèmes observés pour les alvéo-dentales.

- Les palatales :
L’examen des géminées palatales nous a posé le problème de la délimitation de l’implosion. Pour [kk]/, nos tracés acoustiques font apparaître après l’occlusion articulatoire des traces de bruit en moyenne fréquences. Ce bruit ne peut être interprété que comme l’indication d’un contact occlusif insuffisamment fermé. L’évolution générale des appuis de la langue au palais ne permet pas d’interpréter autrement les relâchements occasionnels d’un ou deux contacts au centre.

Il n’est pas possible de mettre en évidence le fléchissement de l’effort articulatoire au milieu de la tenue auquel se succéderait une nouvelle progression des appuis linguo-palataux indiquant une deuxième articulation. Il semble donc bien que pour les palatales, la nature de l’articulateur principal (i.e. le dos de la langue) soit responsable des traces de bruit.

La remarque précédente s’applique sans réserve aux consonnes dévoisées.

- Les alvéo-dentales :
Pour les consonnes alvéo-dentales /t/ et les groupes /dh/, on constate une boucle de corrélation entre les événements articulatoires et les événements acoustiques.

Le schéma général d’organisation des gestes
articulatoires fait apparaître l'existence d'un seul mouvement lingual consistant dans une progression régulière des appuis linguo-palatins pendant la tenue consonantique. (cf. fig. 1)

Toutefois, dans le détail, les phénomènes apparentes en ne plus complexes. Nous avons pu observer à plusieurs reprises une très légère désocclusion accompagnée de bruit. Il s'agit d'un phénomène très bref dont on peut se demander s'il joue un rôle dans la perception d'une consonne simple ou double. Des expériences de Repp [7] que nous comptons reproduire indiquent que non. Il nous semble que l'on doive interpréter la trace très brève de bruit comme l'indication d'un déplacement de la masse linguale sous l'effet d'une grande force d'articulation. Il nous semble, en effet, que s'il y avait un réarticulation, ce phénomène aurait dû aussi être observé sur les courbes des appuis linguo-palatins, et se manifester par exemple par un échellement de la tension musculaire au passage de la première à la deuxième partie de la géménée: ce qui n'a pas été le cas.

4.2. Les assimilations de voûtement
Les traits de source ont souvent été considérés comme des traits redondants des traits de force d'articulation. Une consonne sonore serait forte tandis qu'une consonne sourde serait faible. Il n'est pas possible de distinguer sur ce critère les groupes de palatales homogènes. En effet, la mesure de l'étendue de l'appui linguo-palatin s'il constitue un des moyens d'évaluer en général la force articulaire, est difficile à interpréter pour cette classe de consonnes en raison des déformations possibles des contacts de la langue en dehors des limites de la plaque palatine. Par contre, pour les alvéo-dentales, cette mesure se prête mieux à des comparaisons. Il apparaît que les groupes de sources assimilées sours sont caractérisées par un contact plus étendu de la langue au palais que les groupes de sources correspondantes. Par anticipation de l'articulation consonantique, on constate généralement un phénomène d'assemblage régressif du voûtement, ce qui conduit à un voûtement ou un assouplissement total de la géménée. Les groupes de consonnes ainsi assimilées (sources ou sourdes) possèdent une durée d'occlusion semblable à celle des groupes de consonnes du même mode.

4.3. La force d'articulation
Le nombre d'électrodes touchées permet d'estimer la force avec laquelle une consonne est articulée. On ne constate pas de différence significative entre le nombre de contacts pour les consonnes simples et les consonnes géménées. Seule la durée de la tenue permet de les distinguer.

4.4. L'explosion
On ne constate pas de différence significative entre la durée d'explosion des simples et des géménées. L'explosion des consonnes géménées est de durée égale ou plus courte que celle des consonnes simples. (cf. fig. 2)

5. Influence du débit
L'influence du débit est testée par Wocjik [8] qui trouve que le débit rapide entraîne une hypo-articulation n'a pas été constaté ni au niveau des consonnes simples, ni au niveau des consonnes géménées. On remarque simplement une durée de tenue plus courte et débit rapide. Cependant, les consonnes simples sont moins affectées que les géménées, indépendamment du mode et du lieu d'articulation. (cf. fig. 3)

Il n'y a pas d'influence du débit sur la durée d'explosion. (cf. fig. 2)

5. CONCLUSION
Les consonnes homogènes "géménées" se comportent comme une consonne seule, dont elles ne diffèrent que par la différence de durée de leur tenue ouclusive. On ne constate pas de trace de réarticulation. La force d'articulation se présente pas de différence significative entre les consonnes simples et les pseudo-géménées. Le débit rapide a un effet plus important sur les consonnes "géménées" que sur les simples.

En ce qui concerne l'explosion, pas de différence de durée significative entre les trois groupes consonantiques étudiés. Le débit n'a pas d'influence sur la durée de l'explosion.

6. REMERCIEMENTS
Ce travail a été réalisé dans le cadre du projet ESPRIT II/BRA 33600.

7. REFERENCES BIBLIOGRAPHIQUES
CONSONANT CLUSTERS AND THEIR CONNECTION WITH THE MORPHOLOGICAL STRUCTURE OF THE KAZAKH WORD

S. Meirbekova

Institute of Foreign Languages, Alma-Ata, USSR

ABSTRACT
Consonant clusters are investigated according to the data presented in the dictionary depending on the availability or absence of morphological boundaries between them. From the point of view of morphemic analysis, there may be three syntagmatic types: intramorphic, intermorphic, and mixed. Peculiarities of consonant clusters in the text in relation to the morphemes are defined as well. Analysis of the data of the dictionary and the text revealed features of similarity and difference.

1. MORPHOLOGICAL ANALYSIS OF THE CONSONANT CLUSTERS ACCORDING TO THE DATA OF THE DICTIONARY
The importance of such a linguistic unit as a morpheme in the syntagmatic analysis of the sound system of a language is obvious. Analysis of consonant clusters in the word, depending on the morphemic boundaries is very important, as such an investigation is connected with the lexicogrammatical aspect of a language.

A Kazakh language is one of agglutinative languages and the morphological structure of words is implicit.

Consonant clusters occur in medial and final positions of words. Two and three-member clusters occur in medial position, and in final position only two-member clusters are used.

-Medial two-member consonant clusters are found in four morphological positions: 1) in the root, 2) on the boundary of the root and the suffix, 3) in the suffix, 4) on the boundary of suffixes.

-It is supposed that there is a certain attraction of consonant clusters to their positions in morphemes or on the morphemic boundaries. 27 consonant clusters are intramorphic, 15 intermorphic, 115 mixed. Intramorphic consonant clusters occur only in the root and are of low frequency. Their frequency is 66.

Frequency of intermorphic clusters is 295. Consonant clusters on the boundary of the root and the suffix are more frequent than on the boundary of suffixes. Consonant clusters of the mixed type are the most characteristic in lexical units of the Kazakh language, because their frequency is 1222.

-Medial three-member consonant clusters are presented in three morphological positions: 1) in the boundary of three suffixes, 2) on the boundary of the root and two suffixes, 3) on the boundary of two consonants of the root and the suffix. Therefore medial three-member consonant clusters are intermorphic. They are more characteristic on the boundary of two consonants of the root and the suffix than on the boundary of the root and two suffixes, since their frequency is 104, 94, 54 respectively.

-Final two-member consonant clusters may be intramorphic and mixed. They occur in two morphological positions: 1) in the root, 2) on the boundary of suffixes. They are characteristic in the root of the word, where their frequency is 127 and they are not characteristic on the boundary of suffixes where their frequency is 2.

2. MORPHOLOGICAL ANALYSIS OF THE CONSONANT CLUSTERS ACCORDING TO THE DATA OF THE TEXT
According to the data of the text, medial two-member consonant clusters occur in 3 morphological positions: 1) in the root, 2) on the boundary of the root and the suffix, 3) on the boundary of suffixes. Consistent clusters on the boundary of the root and the ending are more frequent than on the boundary of suffixes. Consonant clusters of the mixed type are the most characteristic in lexical units of the Kazakh language, because their frequency is 1222.

-Medial three-member consonant clusters are presented in five morphological positions: 1) in the boundary of three suffixes, 2) on the boundary of the root and two suffixes, 3) on the boundary of two consonants of the root and the suffix, 4) on the boundary of the root and the ending, 5) on the boundary of the root and the suffix, 6) on the boundary of suffixes.

Frequency of consonant clusters is 135. 860. Frequency of text consonant clusters is 1. 633. The frequency of frequency of text consonant clusters is 1. 633.

Such a great number of morphological positions is explained by the fact, that words in texts are given in different grammatical forms, while words in dictionaries are given in their initial forms.

-23 consonant clusters are intramorphic, 3 intermorphic, and 76 mixed. Their frequency is 135, 860 and 4558 respectively. In intermorphic clusters occur only in the root of words. The most frequent intermorphic consonant clusters are observed on the boundary of the root and the suffix in the ending and on the boundary of endings. Consonant clusters of the mixed type are the most frequent on the boundary of the root and the suffix in the root, on the boundary of suffixes and on the boundary of the root and the ending. Average frequency of consonant clusters is observed on the boundary of suffixes, on the boundary of the suffix and the ending, and on the boundary of endings. Consonant clusters in the ending are more frequent on the boundary of the root and the suffix in the root, on the boundary of suffixes and on the boundary of the root and the ending. Average frequency of consonant clusters is observed on the boundary of the suffix and the ending, on the boundary of endings, and on the boundary of the word. Frequency of consonant clusters is low in the suffix and on the boundary of the ending and the suffix.

-Medial three-member consonant clusters are presented in five morphological positions: 1) in the boundary of three suffixes, 2) on the boundary of the root and two suffixes, 3) on the boundary of two consonants of the root and the suffix, 4) on the boundary of the root and the ending, 5) on the boundary of the root and the suffix, 6) on the boundary of suffixes.
In the rest three morphological positions frequency is low. Intramorphemic three member consonant clusters are the most frequent in the text, mixed consonant clusters are less frequent. Intramorphemic clusters occur very rarely. Their frequency is 9%, 8%, 7% respectively.

- Final two member consonant clusters occur in two morphological positions: 1) in the root, 2) on the boundary of suffixes. Their frequency is 81% and 2% respectively. These consonant clusters may be intramorphemic and mixed. Their percentage is 8% and 19%.

3. CONCLUSIONS

- As a result of the comparison of morphological analysis of consonant clusters according to the data of the dictionary with the data of the text there may be the following conclusions:
  1. The quantity of the morphological positions of medial two and three member consonant clusters according to the text exceeds the quantity of morphological positions according to the dictionary.
  2. Final two member consonant clusters are presented in two identical morphological positions both according to the dictionary and according to the text.
  3. According to the data of the text and the dictionary medial two member consonant clusters may be intramorphemic, intermorphemic and mixed.
  4. Medial three member consonant clusters according to the text may be of 3 syntagmatic types, while according to the dictionary they are only intermorphemic.
  5. According to the dictionary and the text the most characteristic are medial two member consonant clusters of the mixed type, less characteristic are intramorphemic and mixed.
  6. Both according to the dictionary and the text the suffixes, the clusters are more probable on the boundary of the root and the suffix than on the boundary of the root and the root, than in the root of the word, in the root they are preferable than on the boundary of suffixes, and on the boundary of suffixes the clusters are used more widely than in the suffix of the word.
  7. Medial three member consonant clusters are productive on the boundary of the root and the suffix both according to the dictionary and the text and non-productive on the boundary of the root and two suffixes. According to the dictionary the boundary of these suffixes is characterised by high frequency, while according to the text that position is characterised by low frequency.
  8. Final two member consonant clusters are frequent in the root and they are of low frequency on the boundary of suffixes both according to the dictionary and the text.
UNDERSTANDING "HM", "MHM", "MMH"

Stefan Werner
University of Joensuu, Finland

ABSTRACT
Various kinds of hm-like utterances occur frequently in everyday discourse. This paper presents an examination of forms and functions in a subset of German hms: hm uttered as reply or reaction to a question. Subjects' ratings of stimuli on a meaning scale from 'negative' to 'affirmative' yielded a clear functional classification. Subsequent phonetic analysis revealed strong correlations with syllable structure and fundamental frequency variation.

1. INTRODUCTION
Sounds transcribable as "hm", "mhm", "uhuh" and so on - henceforth generically called hm - can be - among other possibilities - a sign of listening, understanding, agreement or disagreement, hesitation, a request to repeat a phrase, an announcement of another speech act, an answer to a question.

But in spite of the obvious importance of hm, it has not yet received too much attention among phoneticians or even linguists (one noticeable exception for German is Ehlisch's discourse-analytically motivated phonetic classification in [1]). My study introduces a first set of acoustic features in German hm that apparently not only modify or differentiate meaning, but suffice to produce it, at least in the semantically limited context used for the experiment.

2. TEST DESIGN
23 test subjects, all of them native speakers of German were asked to rate the meanings of different realizations of hm, presented in random order as the answers to simple yes/no questions, on a scale from 1, 'clearly negative', to 4, 'clearly affirmative' (with the possibility to omit the answer in case of ambiguity). 21 hm stimuli out of 70 recordings had been selected by a jury of two native speakers as a sufficiently large and representative collection. Three different questions were each used twice with every stimulus.

3. TEST RESULTS
Since each subject rated all 21 hm types six times, the ideal ordinate scale for these settings comprises not just four, but 216=126 ranks. Figure 1 shows the sorted mean ranks of all hm types and their standard deviations (the use of these ratio scale statistics for this diagram being justified by the fact, that mode and median in all cases are extremely close to the arithmetic mean and stray values are rare.)

The division into four groups seems obvious, but let us first of all strengthen the case for a clear distinction between hm as a negative and hm as an affirmative answer: figure 2 presents the respective shares of ratings falling below and above the theoretical division line between ranks 63 and 64.

The separation is, in fact, evident. The same point can be made by means of a cluster analysis: a Ward dendrogram exhibits an extreme increase in heterogeneity between the clusters of hm types 1 to 11 and 12 to 21. In addition, there were no missing observations, i.e. ambiguous cases, at all.

On a less significant level, also the subdivisions suggested by figure 1 can be verified with different methods; cluster analysis supports the existence of four groups as well as figure 3 does.

4. PHONETIC ANALYSIS
In order to find acoustic predictors for the negative versus affirmative meaning of a hm utterance (or even for its membership in one of the subclasses), each stimulus' duration, intensity, F0 and spectrums were examined. The main results are:
- the clue to the functional dichotomy is provided by two clearly distinct types of fundamental frequency contours
- the subdivision is related to the existence of one versus two intensity peaks (monosyllabic vs. bisyllabic hm)
- among bisyllabic hms, there is a second criterion for differentiation: the second syllable of a negative hm starts with a glottal stop, an affirmative one has in the same place a /h/.

Figure 4 shows two prototypical F0 contours. This opposition of curvy and flat can be found not only in German, but presumably in a large
least in certain contexts, convey meaning the same way 'normal' words do: by utilizing phonetic features alone.

A link between experimentally established meaning classes and phonetic characteristics was presented. Future research should take into account a wider range of hm types and contexts from various languages.

6. REFERENCES
[1] EHLICH, K. (1986), "Interjektio-
nen" Tübingen: Niemeyer
speakers' perceptions of English
"non-lexical" intonation signals", 
Language Learning 33(1)
[3] WERNER, S., IKONEN, U., 
NIEMI, J. (1984), "Observations on 
Finnish discourse-interjections", in 
Ikonen, U., Tikka, T. (eds.), Papers 
from the Twelfth Meeting of Finnish 
Phoneticians - Joensuu 1984, Joensuu: 
University Press

5. CONCLUSION
hm utterances in German can, at

Fig. 5

least in certain contexts, convey meaning the same way 'normal' words do: by utilizing phonetic features alone.

A link between experimentally established meaning classes and phonetic characteristics was presented. Future research should take into account a wider range of hm types and contexts from various languages.

6. REFERENCES
[1] EHLICH, K. (1986), "Interjektio-
nen" Tübingen: Niemeyer
speakers' perceptions of English
"non-lexical" intonation signals", 
Language Learning 33(1)
[3] WERNER, S., IKONEN, U., 
NIEMI, J. (1984), "Observations on 
Finnish discourse-interjections", in 
Ikonen, U., Tikka, T. (eds.), Papers 
from the Twelfth Meeting of Finnish 
Phoneticians - Joensuu 1984, Joensuu: 
University Press
Age effect on acquisition of non-native phonemes:  
perception of English /r/ and /l/ for native speakers of Japanese

Reiko A. Yamada, & Yoh'ichi Tohkura
ATR Auditory and Visual Perception Research Laboratories

ABSTRACT

This study investigates the age effect on acquisition of American English (AE) /r/ and /l/ perception by native speakers of Japanese who have once been exposed to the AE speaking environment. A perceptual experiment designed to test the ability to identify naturally spoken /r/ and /l/ and determine perceptual cues when identifying those phonemes using synthesized stimuli was performed for native AE subjects and native Japanese subjects with and without the experience of living in the U.S. The results show that some of the Japanese subjects who had resided in the U.S. acquired /r/ and /l/ perception and that acquiring capability of acquiring decreases from 7 to 13 years of age.

1. INTRODUCTION

Many studies have revealed that phoneme perception is modified by the linguistic environment. The perception of American English (AE) /r/ and /l/ sounds for Japanese speakers is one of the strongest pieces of evidence that this is so. In the phonological system of Japanese, the AE /r/ and /l/ contrast is not distinctive, and neither AE /r/ and /l/ resemble any Japanese phonemes. Thus, most Japanese speakers have considerable difficulty in acquiring /r/ and /l/ contrast even though they start learning English in junior high school at about age 12.

Previous cross-linguistic studies using a synthetic /r/-/l/ stimulus series revealed that native speakers of Japanese had difficulties in perceptually differentiating these two phonemes, and that they perceive the synthetic /r/-/l/ series continuously, even though native AE speakers perceive them categorically (e.g. [5, 6, 7, 9]). Furthermore, the perceptual cue for distinguishing /r/ from /l/ is different between AE speakers and Japanese speakers: AE speakers use F3 frequency as a predominant cue, and Japanese speakers use both F2 and F3 frequencies (12). The effect of age on the /r/-/l/ acquisition. However, further control of the starting age and period of exposure are needed to understand the nature of acquisition process. Furthermore, the age of the subjects during participation in the experiment should also be controlled (e.g. it varied from 3 to 45 years of age in [11]), because the performance of children and adults may be expected to differ considerably.

This paper investigates the age effect on acquisition of AE /r/ and /l/ phonemes for native adults of Japanese by controlling the starting age and period of exposure to the AE speaking environment more precisely than previous studies. To determine the precise perceptual mode of the subjects, the identification tests not only of naturally spoken stimuli, which were designed to see overall identification ability, but also of synthesized stimuli, which were designed to investigate the perceptual cue, were performed. Furthermore, in this paper, the /w/-/l/ phoneme is considered in addition to /r/ and /l/, because Japanese listeners are often used to listen to stimuli and confuse some of the /r/ and /l/ sounds with /w/ (13).

2. STIMULI

Synthesized /rai/-/lai/ series generated by Klatt's cascade formant synthesizer, and naturally spoken stimuli were used. Figure 1 provides a schematic spectrographic representation of the initial CV portion. /rai/-/lai/, for the synthesized stimuli. The acoustic parameters for idealized “right” and “left” were derived from the naturally spoken /rai/ and /lai/ stimuli by a native male speaker of AE. When generating the stimuli, three acoustic parameters, F2 and F3 onset frequencies and F1 transition, were varied. To construct the stimuli on F2-F3 plane, a variety of F2 and F3 onset frequency combinations were used. The F2 and F3 onset frequencies were varied independently from 800 Hz to 1400 Hz in 200 Hz steps, and 1200 Hz to 3000 Hz in 200 Hz steps, respectively. There were 12 combinations in total, including some contradictory combinations in which the F2 frequency was equal to or higher than the F3 frequency. F1 transition duration was varied from 20 ms to 16 ms in 20 ms steps as the F3 onset frequency was varied from 1200 Hz to 3000 Hz. In all synthesized stimuli, the acoustic parameters for the vowel part /ai/ were common, and the duration of the /rai/ was fixed at 360 ms. The stimuli were synthesized and reproduced through 16-bit digital audio conversion at a sampling frequency of 20 kHz and low-pass filtering with a cutoff frequency of 10 kHz. Several experimental sessions (i.e. with different stimulus randomizations) were recorded on a digital audio tape using a DAT recorder, SONY DTC-1000ES. Each session consisted of one block of ten trials and one block of six trials. Other conditions were identical to the identification tests of the synthesized stimuli.

3. SUBJECTS

One hundred and twenty native speakers of Japanese who have never lived abroad (Group J), 109 native speakers of Japanese who have resided in the U.S. (Group JE), and 9 native speakers of AE (Group A) served as subjects. Criterion for participation in the experiment was Group JE subjects was to fulfill all the following conditions: (1) native speaker of Japanese, (2) born and raised in Japan, (3) no exposure to American English before age 12, and (4) have lived in the United States for at least two years before the experiment. Group JE subjects were divided into two groups based on how long they had lived in the United States: Group J (36 subjects) and Group JE (73 subjects). Group J subjects were native speakers of Japanese who had lived in the United States for at least two years before the experiment. Group JE subjects were native speakers of Japanese who had lived in the United States for at least two years before the experiment. Each subject participated in a single experimental session.

Figure 1. Schematic representation of frequency trajectories of F1 to F3 for the synthesized stimuli.
(2) had once lived on the U.S. mainland for more than 1 year, (3) had never lived in a foreign country other than the U.S., (4) speaks AE all the time at school, preschool or kindergarten, or in business, (5) goes to school or conducts business under condition (4) at least 5 days a week, (6) received no special training for speaking AE in Japan. The start of their residence in the U.S. can roughly be thought to coincide with the start of their exposure to the AE speaking environment because English education in Japanese high schools is biased toward grammar, reading, and writing, and is mainly conducted by Japanese teachers. The age of the subjects in Group J was 19, on average, and ranged from 15 to 23, that in Group JE was 20 on average, and ranged from 13 to 40, and that in Group A was 25 on average, and ranged from 20 to 41. All the subjects reported no history of hearing or speaking disorder.

4. PROCEDURE

Each listener participated in two sessions of identification tests for synthesized stimuli, and one session of identification test for naturally spoken stimuli. In these tests, listeners were instructed to identify the word initial consonant, and to make a forced choice among the given categories regardless of the frequency of occurrence for each category through an entire session by checking a corresponding response category on an answer sheet. In the identification test for naturally spoken stimuli, listeners were told that there might exist unfamiliar or meaningless words, but they should only identify the initial consonant.

5. RESULTS

After the identification rates for each stimulus were calculated, the values Cs and Cn were obtained as perceptual ability scores of synthesized stimuli and that of naturally spoken stimuli, respectively. As AE listeners identify the stimulus whose F3 onset frequencies were higher than 2000Hz as /l/ and those which were lower as /r/ (Yamada & Togihara, 1990), the Cs represents the averaged response rates of /l/ for the stimuli whose F3 onset frequencies were equal to or higher than 2000Hz and /r/ for the other stimuli (0 ≤ Cs ≤ 1). The Cn is the averaged correct response rates across all the naturally spoken stimuli (0 ≤ Cn ≤ 1).

The averaged Cs across Group A subjects was .91, and ranged from .75 to 1.00, that across Group J subjects was .48, and ranged from 21 to .76, and that across Group JE subjects was .74, and ranged from .32 to .99. The averaged Cn across Group A was 1.00, that across Group J was .67, and ranged from .44 to .95, and that across Group JE was .87, and ranged from .55 to 1.00. In the histograms of both Cs and Cn values, two peaks were observed in Group JE, even though only one peak was observed in Group A and J.

The Group JE subjects were divided into two groups according to their Cs and Cn values as follows: acquired group (subjects whose Cs and Cn values are: 0.75 ≤ Cs, and 0.90 ≤ Cn), and non-acquired group (the other subjects). In order to observe the correlation between the acquisition performance and the age of exposure to the AE speaking environment, the probabilities of acquired group subjects among subjects who have started living in the U.S. at the same age were calculated. JE subjects were classified into groups according to their living periods, and the following four groups were represented in Figure 2: subjects living in the U.S. for 1 year, 2 - 3 years, 4 - 6 years, and 6 + years. As the living conditions (starting age and period of residence) are not fully controlled, and the number of subjects for each data point was not insufficient, we plotted the moving averages under the following conditions: the average age upon taking up residence was 5 years, and the shift period was 1 year. Age noticeably age affected acquisition performance. The acquisition probability decreased rapidly from 7 to 13 years of age. This result is especially obvious in the 2 - 3 years old, in which living conditions are better controlled than in the other groups. Eleven subjects have resided in the U.S. more than 8 years, only one of them, who have resided in U.S. for 8 years from 25 years old, failed to acquire /r/ perception.

6. DISCUSSION

Showing that result that the acquisition probability decreased with age was consistent with many previous studies of phoneme acquisition (e.g. [1, 2, 3, 4, 10]). The age when the capability of acquisition decreases in the present study was also similar to results of previous studies on second language production (e.g. [8, 11]). The relationship between acquisition of perception and that of production is of great interest. The productions of /r/ and /l/ phonemes for all the subjects in the present experiment were recorded after the perception test. We also plan to analyze the production characteristics of the present subjects and to study the relationship between acquisition of perception and that of production. In addition, further efforts to obtain data from Japanese subjects with greater variations in living conditions are required.

7. REFERENCES

THE REDUPLICATIVE BABBLING OF FRENCH- AND ENGLISH-LEARNING INFANTS: EVIDENCE FOR LANGUAGE-SPECIFIC RHYTHMIC INFLUENCES

Andrea G. Levitt1 and Qi Wang1

Haskins Laboratories, New Haven, CT, Wellesley College, Wellesley, MA, University of Connecticut, USA

ABSTRACT

The reduplicative babbling of five French- and five English-learning infants produced between the ages of five and thirteen months was examined for evidence of language-specific rhythmic patterns. The babbling of the French infants showed a significantly greater percentage of final-syllable lengthening than that of the American infants. The French babbling showed more regularly timed nonfinal syllables than that of the Americans, although only in the later stage of the infants' reduplicative babbling. The French infants also produced significantly more reduplicative babbles that were four or more syllables in length.

1. INTRODUCTION

Jakobson's [6] famous proposal of discontinuity between babbling and early that were not found much support in current research on child language acquisition. Instead, many have found evidence of continuity between babbling and early speech (e.g. [7]). The child's babbling seems to "drift" [4] in the direction of phonetic characteristics of the ambient language.

The question of how early the child's productions reflect the segmental properties of the native language has been much debated, with some finding evidence for such differences during the early first year of life (e.g. [2]) while others do not (e.g. [9]). Very little attention has been devoted to the early stages of prosodic development, although some have suggested (e.g. [3,10]) that infants may begin to imitate the prosodic patterns of their language earlier than they imitate the segments. In a recent investigation [15], we found evidence for language-specific effects in the F0 contours of the reduplicative babbling of French- and English-learning infants. In the present investigation we extended our study to the rhythmic properties of those reduplicative babbling, in particular phrase-final lengthening, the timing of individual syllables within each utterance, and the number of syllables per utterance. Both French and English exhibit final syllable lengthening (breath-group final lengthening in French), but because French nonfinal syllables are not typically lengthened due to word stress, final-syllable lengthening is a more salient feature of French, which is "taller timed," according to Wenk and Wioland [4]. There has been some indication that French and American infants may develop final-syllable lengthening fairly early on. In examining the babbling of a group of French-learning infants, Boysson-Bardies [7] found that final syllables were longer on average than nonfinal syllables, from the age of eight months on, although this difference did not become significant until the children were 16 months old. Oller and Smith [12], in examining the babbling of six or seven infants ranging in age from 8 to 12 months, found evidence for such lengthening in the babbling of some but not all of their infants. However, it is not clear whether the onset of such lengthening might differ between the two groups, our study looks at French and English babbling both longitudinally and cross-linguistically.

In terms of nonfinal syllable timing, French has been classified as syllable-timed (e.g. [13], but cf. [14]), with a rhythmic structure known as isosyllabicity, which is characterized by nonfinal syllables generally equal in length. Because word stress in English tends to lengthen nonfinal stressed syllables, English does not exhibit isosyllabicity. If French nonfinal syllable timing has an effect on the infant's productions, then we would expect the French infants to exhibit more regularly-timed nonfinal syllables.

Finally, in keeping with the possibility for phonetically-based breath groups in French to contain as many as four to six syllables, whereas intervals between stressed syllables in English rarely contain more than four syllables, we expected that our French infants might produce longer reduplicative utterances than our American infants. Indeed, Boysson-Bardies [16] reported a similar effect of utterance length for somewhat older children.

2. PROCEDURE

2.1 Subjects

The babbling of five English-learning infants (three male and two female) and five French-learning infants (four male and one female) was recorded weekly by their parents at home. The French-learning infants were recorded in Paris and the English-learning infants were recorded in the southeastern United States. The average age of the infants at the first recording used was 7.3 and the last was 11.1 months (ranging from 5 to 13 months).

2.2 Method

The infants were recorded on cassette tape using high quality microphones. Home recording sessions lasted between 10 and 20 minutes. Parents were instructed to choose a time when their infant was alert and unlikely to cry. They could elicit babbling by talking and gesturing, but they were told to be sure to stop speaking as soon as the infant began vocalizing. The microphone was to be held about 20 cm from the baby. The parents identified each individual tape by recording the date at the beginning of each session. A comment sheet was also filled out for each tape and included the date, time, and situation (e.g. "in bath") of each recording.

Each tape was transcribed, and all infant vocalizations (except for squeals, grunts, emotive sounds, and vegetative noises) were digitized at 16 kHz via the Haskins Laboratories PCM system [16]. The vocalizations were divided into utterances, or breath groups, which were defined as a sequence of syllables that were separated from other utterances by at least 750 ms of silence and which contained no silent periods longer than 450 ms in length. From the phonetically-stressed syllables in each utterance, we selected all the reduplicative babbles according to our transcripts. Using these criteria, we obtained 208 reduplicative utterances, approximately half (102) from the English-learning children and half (106) from the French-learning infants.

Reduplicative babbles consist of two or more repetitions of the same syllable, which in the case of our ten infants, were all open CV syllables. Because phonetic segments are of inherently different lengths (for example, syllables are typically longer than stops), we analyzed only reduplicative babbles, where all the consonants and vowels in a single utterance are the same, in order to eliminate syllable duration variations due to inherent differences in segment length.

The duration of each syllable was measured using a wave form editing and display program. A conservative criterion for measuring syllable length was adopted, such that duration measurements only included the visibly voiced portion of each syllable. This criterion was adopted because the home recording environments were occasionally noisy, and the noise could serve to obscure, in some cases but not in others, the breath release of certain syllables. Although nonfinal syllable lengths could be considered to extend to the onset of the following syllable, such an alternative measure was not available for final syllables, inasmuch as the distance between nonfinal and final syllable lengths problematic. Thus, in order to avoid such difficulties, breath releases and intersyllabic spaces were not included in the syllable measurements.
3. RESULTS
We measured final syllable lengthening by comparing the length of the final syllable of each reduplicative utterance to that of the penultimate syllable. For each infant, we calculated the percentage of utterances showing final syllable lengthening. The French infants showed final syllable lengthening in 63% of the utterances on average, whereas the American infants showed final syllable lengthening in 42% of their utterances. This difference was significant [t(8)=2.37, p=.027, one-tailed].

In order to see whether this pattern was evident throughout the period during which reduplicative babbling was detected for each child, we divided each infant’s utterances into two groups. The first group, the “early” stage of reduplicative utterances, was produced during the first half of the time period and the second group of “late” reduplicative utterances was produced in the second half of the time period. We again calculated the mean percentage of final syllable lengthening for each infant during the early and during the late period. The results of an ANOVA with repeated measures indicated again an overall group effect of language background [F(1,8)=6.379, p=.0227, one-tailed], but no effect of early vs. late utterances and no interaction of language background and early vs. late utterances.

We measured iso-activity, i.e., the relatively regular timing of nonfinal syllables within each utterance, by calculating the standard deviation of the nonfinal syllables for each utterance and determining the mean standard deviation for each infant. Although the French infants did show lower standard deviations on average (54.5), indicating more regularly timed utterances, than the English (65.4), the difference was not significant.

In order to see whether there was a significant shift in this tendency over the period during which reduplicative babbling was detected for each child, we again calculated the utterances by time period into two groups, the early and the late. The mean standard deviation was again calculated for each infant during the early period and the late period of babbling for each infant. An ANOVA with repeated measures (early vs. late percent of long utterances) was conducted on the results. Again, there was a significant main effect of language background [F(1,8)=6.379, p=.0355], but there was no significant main effect of early vs. late percent of long utterances nor any significant interaction of language background and early vs. late percent of short utterances.

4. DISCUSSION
We found acoustic evidence for language-specific rhythmic effects in the reduplicative babbling of French and English infants. In particular, French infants produced a higher percentage of final-syllable lengthening and of utterances four or more syllables in length. In addition, French infants produced more regularized nonfinal syllables, although only in the later stage of their reduplicative babbles. However, whereas our study of the FO properties of our infants’ reduplicative babbles [15] revealed both acoustic and perceptual effects, the rhythmic differences that we have discerned here do not appear to be sufficiently robust to be detectable by adult listeners. Nonetheless, just as Macken and Barton [11], through acoustic analysis, discovered that children learning the voicing distinction in English went through a stage during which they produced the contrast in a manner that was not perceptible to adults, we believe that our results represent a similar stage in the acquisition of prosody. Indeed, as Allen [1] has shown, French children exhibit many of the prosodic characteristics of their language in a more robust fashion by two years of age.

Thus, our results, along with those of Boysson-Bardies and her colleagues [3] suggesting that the babbling of infants younger than one year of age may reveal language-specific vocalic and prosodic influences when analyzed acoustically.

5. ACKNOWLEDGMENT
This work was supported by NIDCD Grant DC-00403 to Catherine Best.

6. REFERENCES
THE UNINTelligibility OF SPEECH TO CHILDREN: EFFECTS OF REFERENT AVAILABILITY

E. G. Bard and A. H. Anderson

Human Communication Research Centre,
Department of Linguistics, Centre for Cognitive Science, University of Edinburgh
Department of Psychology, University of Glasgow

ABSTRACT

Speech addressed to children is supposed to be helpful in several ways, but redundant words in speech to adults tend to lose intelligibility [7,8]. Word tokens extracted from the spontaneous speech of the parents of 22- to 36-month-old children and presented in isolation to adult listeners lose intelligibility when the words are redundant in two senses: they occur in repetitions of an utterance (Experiment 1) or they refer to an entity which is physically present at the time of speaking (Experiment 2). These findings help to explain why word tokens randomly selected from speech to young children are less intelligible than those from speech to adults [2]. Because these tokens are difficult to recognize, they appear to induce child listeners to rely on the word's extra-linguistic context during the recognition process [1], much as adults are induced to rely on discourse context [3,6].

1. INTRODUCTION

Children perform a remarkable bootstrapping operation when they are not yet familiar with syntax and vocabulary by listening to running speech. Word tokens in spontaneous speech differ so greatly from their citation forms that they have a 50% chance of being recognized in isolation by adult listeners who share the speaker's vocabulary [10]. Given that the child's interpretation of linguistic context may be too incomplete to aid word recognition in all cases, categorization as arosional tokens as belonging to a particular word type or learning more about the structure of a language from strings of such tokens must be especially difficult.

The perceptual task might be simplified if parents ambiguously spoke more clearly to children than to adults, but on the contrary, words randomly selected from parents' speech to children (hereafter "A-C speech") aged 22 to 36 months proved significantly less intelligible out of context than words from the same parents' speech to an adult (hereafter "A-A speech") [2]. Alternatively, the well attested redundancy of speech to small children [9] may make their task easier. Words are more predictable from their sentence context in A-C speech than in A-A speech [2]. Utterances to children are more often partly or completely repeated [9,11].

A-C speech is also more supported by physical context, since it refers almost exclusively to objects and situations which are available to the child's senses at the time [9]. Perhaps some combination of the surrounding situation, earlier occurrences of the same utterance, and the physical presence of referent objects can be exploited by the child.

In A-A speech, however, more redundant word tokens, both those more predictable from sentence context [7,8], and those referring repeatedly to the same entity [4-6], are shorter and less intelligible when isolated than their less redundant counterparts. If the effect applies for all kinds of redundancy, then words naming salient visible objects may also be less clear. In A-C speech, increased intelligibility from sentence context has been found to correlate with lowered word intelligibility [2]. This paper asks whether intelligibility also falls when A-C words refer to just mentioned entities or denote physically present objects.

2. EXPERIMENT 1: REPETITION

Experiment 1 tests the hypothesis that words which are highly identical A-C utterances produced in close succession will be less intelligible than words in the first.

2.1. Method

Corpus. The materials were drawn from 12 45-minute studio-recorded sessions, in which each parent spoke to his or her child and to an experimenter. Both parents of one boy and one girl in each of three age groups (22-24 months, 28-30 months, 34-36 months) participated. After discussing with the parents the family's history and details of the child's contacts and play habits, the experimenter encouraged the parent to help the child play with a standard set of toys so that the child's speech in play might be recorded. The parent later ed the child in conversation about one of his or her own toys which resembled one in the studio. Parent and child were recorded on separate channels of a Revox A77 stereo tape recorder, the parent via a lavelier microphone. Other details will be found in [2].

Tapes were fully transcribed in the standard orthography and all nouns spoken by the parents, except proper names, were classified according to the address and the location of the entity referred to. Present nouns named objects or persons in the studio which were being discussed or acted on by speaker and listener. Present-for-Speaker nouns referred to objects to which the speaker was thereby directing the listener's attention. Absent nouns referred to entities or events not present in the studio. Unclassifiable nouns referred to abstractions and to physical or linguistic entities in which the studio was contained. Table 1 summarizes the different distributions of A-C and A-A nouns among these categories.

Materials and Design. From the speech of each parent to his or her child, 4 pairs of word tokens were chosen. Each pair included two successive co-referential tokens of a single noun which occurred in self-repetitions in the same conversational turn, the second of which either exactly repeated or closely paraphrased the first without altering the noun phrase containing the selected word. Two pairs from each parent were Child-Parent-present words, two Child-Absent.

The selected items were excerpted from their taped contexts electronically and distributed among four groups to give balanced representation of speaker, token, and location. No group contained more than 75% of a parent's material. Each was presented in random order interspersed with materials from Experiment 2. Intensity levels were held constant as far as possible. Each word was preceded by a spoken number and repeated three times at approximately 5 sec intervals.

Subjects and Procedure. Twenty-four native speakers of English from the Edinburgh University community heard stimuli presented monaurally on a Revox A77. They were told that each stimulus was a word taken from conversational speech which they were to identify, by guessing if necessary.

2.2. Results

Figure 1 summarizes the results. The number of letter perfect or fully homophonic identifications of the stimuli showed the expected effect of Token: first tokens were more intelligible than second tokens (57.5% vs. 41.5%: F(1,1) = 11.84, df = 1,2, p < 0.05; F(3) = 3.92, df = 1, 44, p < 0.05, Min F* = 2.94, df = 1, 64, 05 < p < 0.10. Thus, A-C speech shares with A-A speech a tendency to lose in clarity what it gains in repetitiveness [4-6].

3. EXPERIMENT 2: LOCATION

Table 1 illustrates a typical asymmetry between A-A speech, which refers largely to absent entities, and A-C speech, which deals with visible things. Even when first mention of a new present word was already "given" by extralinguistic context. The other location categories may include mentions which introduce new items. If linguistic and extralinguistic contexts work similarly, then Present nouns should resemble co-referential or Given second mention in having relatively low intelligibility [4,6], whereas other categories

Table 1. Distribution of Common Nouns from Speech of Twelve Parents by Addressee and Location of Referent (N = 4013; χ² = 1371, df = 3, p < .0001)

<table>
<thead>
<tr>
<th>ADDRESSER</th>
<th>PRESENT</th>
<th>PRESENT FOR SPEAKER</th>
<th>UNCLASS.</th>
<th>ABSENT</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHILD (%)</td>
<td>1587 (62)</td>
<td>80 (3)</td>
<td>495 (19)</td>
<td>406 (16)</td>
<td>2598</td>
</tr>
<tr>
<td>ADULT (%)</td>
<td>70 (3)</td>
<td>7 (0)</td>
<td>583 (23)</td>
<td>785 (31)</td>
<td>1445</td>
</tr>
</tbody>
</table>

The intelligibility of A-C speech to children is compared with that of A-A speech to adults, showing a significant effect for the location of referent.
will include more intelligible words. The overall intelligibility difference between A-C and A-A speech might be partly due to the typical referent location for each, and should be lost if this factor is controlled.  

3.1. Method

The corpus allowed balanced sampling from each parent only in Child-Recent, Child-Unclassifiable, Child-Absent, Adult-Unclassifiable, and Adult-Absent categories. From each of these, 4 tokens per parent were randomly selected. The 240 word tokens were prepared by the method described earlier and presented with the 96 tokens of Experiment 1 to the same 24 Subjects.

3.2. Results

Figure 2 shows the means for the 5 cells. Among the A-C words, the predicted effect of location was found: nouns with Absent referents were significantly more intelligible (65% correct recognitions) than those with Unclassifiable (43%) or Present referents (49%), while the latter did not differ significantly: one-way ANOVAs for Referent Location gave $F_2 = 29.14, p = .05; F_3 = 2, 132, p < .05; F_{12} = .1, t = 2, 132, p < .05; Min F^m. s., Scheffé tests at $p < .05$.

For words to both Addressees, Unclassifiable nouns were less clear (49% correct) than Absent (62.5%), though the difference was significant only for words spoken to children: $t(56) = 4.08, t^* = 1, 193, p < .05; Scheffé test by Subjects at $p < .05$. Since neither the Addressee effect nor the interaction was significant, there was no intelligibility difference due to Addressee alone.

4. GENERAL DISCUSSION

Sources of redundancy in speech to small children have a price. When a utterance is repeated, its words become less intelligible. When the objects spoken of are present to the senses, the referring nouns are also less intelligible. The effect cannot be attributed to occasional lapses in generally clear speech, for no A-C cell provides significantly clearer word tokens than the A-A cells (e.g., Child-Absent at 65% vs. Adult-Absent at 60%). When redundancy rises, as in second tokens of repeated words (43.5%) or in Child-Recent words (49%), intelligibility falls to below A-A levels. Even the Unclassifiable words patterned like the clearer Absent cells (55%). Whatever the internal breakdown of the Unclassifiable cells may be, they do nothing to maintain high intelligibility for speech to children.

Of the A-C figures, the lower ones must be taken as typical. Although the present corpus is not a random sample of conversation types, it resembles those in other studies [11]: Child-Recent words, which should be relatively unintelligible, predominate, while the clearer Child-Absent words were relatively rare, even when parents were instructed to produce them. Although self-repetition is a common in A-C speech as Present reference, it is certainly more typical here than in adult conversation [11]. Consequently, the differences in intelligibility of large random samples of A-A and A-C speech [2] may have something to do with the tendency of speech to children to provide the expensive forms of redundancy which have been explored here. Certainly the difference is lost when referent location is held constant.

By succumbing to processes which reduce the meaning, children's support is high. Parents seem to be placing their young children at a disadvantage. To see how children might actually profit from these differences in Intelligibility-Consider the uses to which adults put repeated word tokens. Fowler and Fousum [6] have shown that second tokens are better prompts to the recall of words associated in discourse with first tokens than are the first tokens themselves. They propose in Experiment 2 that the reduced second token signal reference to earlier material and so evoke the associated word. Alternatively, the process of recognizing the less intelligible second tokens may rely more heavily on linguistic context, thereby reactivating a representation of that context [13]. To behave like adults, children would have to map less intelligible tokens onto known items while failing to do this for more intelligible words.

Exactly this result was found when three-year-olds were asked to fetch the toys a puppet 'spoke' in a tape-recording of the words excepted from the present corpus [11]. The children were always familiar with the nouns and the toys available, but in one condition they could see the toys as the puppet 'spoke', while in the other the toys were concealed in a box. Like the adult listeners in Experiment 2, the children found originally Absent words easier to recognize (59%) than originally Present words (45%) over all the case regardless of original addresser. Moreover, originally Present words were more readily identified when the toys were visible than when they were hidden (72.5% correct, $N = 17, v = 36; W = 279, r = 1.99, df = 48, p = .05$), whereas originally Absent words were less accurately identified when the toys were visible than when they were hidden (51%, $N = 30, v = 76; W = 279, r = 2.45, df = 44, p = .019$). Since children knew that all toys would be hidden or all would be visible in a given session, word pronunciation did not signal referent location. Instead children appeared to profit from the visible context to decode uninformable Present words, while that context proved a distraction when they attempted to decode more intelligible Absent words. If these children performed in a typical way, then the unintelligibility of A-C speech encourages them to use supporting context in the process of recognizing what has been said to them. It is fortunate that this context is so often pertinent.

This work was supported by an ESRC IRC grant to the University of Edinburgh and by ESRC Project Grant (SSR 9161030) to John Laver and the first author. The authors are grateful for the help provided by the participants in the study. Recent requests should be sent to the first author at the Human Communication Research Centre, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LL, U.K.

5. REFERENCES

ON THE PHONETIC SYSTEM EVOLUTION IN SOME ARCHAIC RUSSIAN DIACRITICS

S. Kniazhev and S. Pogarskaya
Institute of Russian Language,
 Moscow State University,
 Moscow, USSR

ABSTRACT

The paper deals with the problem of the evolution of some archaic Russian dialects. The development of their phonetic system tends to the elimination of the sound oppositions which are phonologically unsupported and not connected closely with the properties of their basis of articulation.

1. INTRODUCTION

Russian dialects display significant differences in their "basis of articulation" (BA), they are manifested in the phonetic, morphological, and acoustic qualities of particular sounds (i.e., the type of vowels labilization, their localization in more front parts of the oral cavity, and the mouth opening in the location and the type of consonantal articulation, etc.). This property particularizes the ability of segments to participate in certain phonological oppositions. Different directions of the evolution of phonological systems and their stability in face of the above characteristics are connected with the deep segmental transformations during the dialects are also determined by the differences in the BA's.

We have chosen for our investigation a group of Archangel dialects from the Voronezh Oblast region since they were examined more than twice since 1967 and 1990. A brief survey of the Archangel dialects' phonetic system is presented below.

2. VOWELS

The vowel system of the dialect is based on three-level triangle of 5 phonemes (a, a, u, o, e) which were found in a stressed syllable before a "hard" consonant and a tendency may be observed for this triangle to be moved in the same consonantal context of unstressed syllables (a, a', u, u', o, o', e, e', A, A', A, A'). However, the system is reduced in other contexts. Thus, in the position between "soft" consonants the system of phonemes (A, A', A, A') is not disturbed by any lexical or morphological parallels may be observed from two-level triangle of 3 phonemes (a, a, u, u, o, o, e, e, A, A, A, A), in which A is a result of all the primary non-neutralized syllables (n'a - n'a', p - p', p' - p', m - m', m' - m', t - t', t' - t', n - n', n' - n', s - s', s' - s', l - l', l' - l', r - r', r' - r', y - y, y' - y', y - y', y' - y', y - y', y' - y'). The same set is found in unstressed syllables (n'a, n'a', p - p', p' - p', m - m', m' - m'). The significance of stress factor in such a system (this kind of vocalic structure is typical for the Russian and Pommern Dialects) is called "karyisme", is weakened as compared with the word "karyisme" in NPM, where the system contains some intermediate slightly labialized sounds to the position before the syllable containing (e) the vowel harmony is possible (u, o, n). To solve this problem the realization of ko, an, and re, may vary within rather wide limits (fe - fe, re - re, respectively). The narrow opening and passive labial articulation result in a weakening or labialization of vowels and cause the vowel variability. The following [j, k'] may be pronounced in words with the position before k and k. Any concrete (but especially k) may cause the following (j, k').

The results presented in this paper show that the solutions of the various dialects' problems may be different. This is due to the fact that the evolution of the vowel system is connected with the following factors: the linguistic context, the type of word and the type of word procliticalization.
palatalized before front vowels and in the word final position: 

<table>
<thead>
<tr>
<th>phoneme</th>
<th>sound</th>
</tr>
</thead>
<tbody>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
<tr>
<td>m'âômo</td>
<td>[m̩]</td>
</tr>
</tbody>
</table>
MODELLING ARTICULATORY INTER-TIMING VARIATION IN A SPEECH RECOGNITION SYSTEM

M. Blomberg
Department of Speech Communication and Music Acoustics, KTH, Stockholm

ABSTRACT
A technique is described that automatically predicts certain cases of pronunciation alternatives. The method utilises the fact that differing realisation of an utterance often depends on variation in the synchrony between two or more simultaneous articulatory gestures. The technique has been implemented in a recognition system based on synthetic generation of reference templates. Varying delay values have been systematically generated by the speech production system. In a pilot experiment, the recogniser behaviour was examined for varying time position of the devoicing of utterance-final vowels.

1. INTRODUCTION
As is well known, the production of speech is a highly complex process that involves the control of several articulatory gestures for realizing the intended sound sequence. Different physiological, psychological and environmental factors contribute in creating variability in the pronunciation of an utterance. It is essential for a recogniser to model this variability in an appropriate way. In this report, we will discuss variability in the time synchronisation between different articulators. We will give an example of this effect, discuss consequences for speech recognition systems and suggest a new method for dealing with this type of variability.

A transition from one phoneme to a following one often involves simultaneous movements of more than one articulator. Details of the acoustic realization depends among other things on timing differences between these articulators. An example of a phoneme boundary where two separate gestures are active is shown in figure 1. The figure shows spectrograms of the Swedish word 'tre', (English: three) spoken by two male speakers. The phonetic transcription is [tre]. The end of the phrase-final vowel changes gradually towards a neutral vowel, similarly for both speakers. The point of devoicing is different, though. Speaker 1 keeps a steady voicing throughout the neutralisation gesture, whilst speaker 2 aspirates the last part of the vowel. An attempt to align the aspirated vowel portion of speaker 2 to the last part of the vowel for speaker 1 would result in a large spectral error. The earlier point of devoicing for speaker 2 causes a great spectral distortion, which will cause problems for most recognition systems.

An early opening of the vocal folds in this example shortens the voiced part of the vowel and prolongs the duration of the prespiratorive segment. Also, the spectral properties of the aspiration will be changed. The tongue will have moved a shorter distance towards its target at the start of aspiration and the spectral shape immediately after the aspiration onset will be quite different compared to the same point in a boundary with a late opening.

Other examples of overlapping articulatory movements are velar opening during vowels before nasals and change of place-of-articulation between adjacent consonants. In the latter case, it often happens that the release from the first consonant precedes the closure of the second one, which will cause a short vocalic segment to occur. If the release occurs after the closure, there will be no such segment.

2. RECOGNITION APPROACH
The acoustic-phonetic decision part of most existing systems are based on spectral matching without taking into consideration the underlying production parameters. Common techniques like dynamic time-warping and Hidden Markov Modelling [5] are able to compensate for a non-linear tempo variation between two utterances but they do not handle timing asynchrony between the production parameters. Stretching and compression of the time scale of the speech signal implies a uniform time transformation of the underlying articulatory parameters. In these systems, the effect will be reflected by large spectral variation at the phoneme boundaries.

A common way to represent pronunciation alternatives is to use context-dependent phonetic rules, formulated by a human phonetic expert [3] and [4]. The rules operate on the input phoneme string and produce several phonetic output strings. However, they mostly use a qualitative description of the effect of varying delay between the articulators. As discussed above, we also need a quantitative description. This requires a description of the phonetic elements in terms of production parameters.

The optional rules can be modified so that they generate a set of pronunciation alternatives at every phoneme boundary. Within the set, the delay between some of the parameters are varied in a systematic fashion. In this way, a quantitative, as well as a qualitative, description of the articulator asynchrony effect is obtained.

The parameter tracking problem can be avoided by using a synthesis technique for producing reference templates, as mentioned in [1]. In this way, knowledge about the behaviour of different parameters can be utilized without the need of tracking them from the speech signal. Instead, their predicted values can be used for generating corresponding frequency spectra and the recognition matching would be performed in the spectral domain.

3. SYSTEM DESCRIPTION
3.1 Recognition System
The recognition system used for this experiment has been described in [1] and [2]. It uses dynamic programming for finding the path through a finite-state network of subphonemic spectra that minimises the spectral distance to a spoken utterance. During the matching of an utterance, an adaptation procedure dynamically normalizes for differences in the voice source excitation function. The subphoneme spectra have not been created by training, as in the majority of current recognition systems, but by a speech production algorithm described below.

3.2 Reference Data Generation
Figure 2 shows a block diagram of the reference template generation component. It is very similar to a speech synthesis system. Its main difference from such a system is that the output consists of spectral sections instead of a speech signal and that the input phonetic description is a network of optional pronunciation alternatives as opposed to a string in the speech synthesis case. The net can describe a single word or the lan-
3.3 Modelling Articulator Asynchrony

For ease of illustration, we will in the following example consider the change of only two parameters; the others are assumed to be constant. This can be displayed in a two-dimensional array. Figure 3 shows a phoneme boundary, where a voicing transition occurs during the tongue movement when going from a vowel to an unvoiced consonant. The tongue movement, described by interpolated formant values, and the voicing transition are represented in the horizontal and the vertical axes, respectively. They are quantised into a low number of steps. The upper and lower horizontal lines represent the tongue movement during voicing and aspiration, respectively. Different delays of voicing offset relative to the start of the tongue movement are represented by vertical lines at varying horizontal positions. The duration of the voicing transition is considered to be short compared to the tongue movement, and therefore there is no need for diagonal connections in the lattice.

![Figure 3. A sub-phoneme lattice representing varying parameter inter-timing in the transition between a vowel and an unvoiced consonant.](image)

**4. PILOT EXPERIMENT**

Instead of running a complete recognition experiment, we studied the chosen method's ability to align the speech signal to a phonetic transcription of the utterance. The two utterances shown in figure 1 were used for testing the method.

To represent the possibility of devoicing of the final vowel, we implemented a subphoneme lattice similar to figure 3, where the consonant in this case is the phrase-end symbol. This symbol is marked in the phoneme library as unvoiced and having neutral formant targets.

The speech signal was analysed by a computer implemented 16 channel Backscale filter bank covering a frequency range from 0.2 to 6 kHz. The frame interval was 10 ms and the integration time was 25.6 ms.

**5. RESULT**

The paths through the network for the two utterances are shown in figure 4. The predicted, interpolated value of the second formant is displayed for every subphoneme. The path for speaker 1 shows a voicing offset at a later stage of formant transition than that of speaker 2. This conforms well with the spectrogram displays in figure 1.

![Figure 4. Results of alignment of the last part of the phrase-final [:] The paths for speakers 1 and 2 are displayed. State values of the second formants are shown.](image)

The accumulated spectral error over the phoneme boundary was also measured. It was compared with the errors using a fixed-delay subphoneme string, having early or late voice offset.

The results in table 1 show that the proposed method works well for both speakers, whereas each of the preset delay values gave low error for one speaker only.

**Table 1. Accumulated spectral error over the final transition interval of the two vowels in figure 1.**

<table>
<thead>
<tr>
<th>Devvoicing</th>
<th>Speaker 1</th>
<th>Speaker 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early</td>
<td>165</td>
<td>110</td>
</tr>
<tr>
<td>Late</td>
<td>133</td>
<td>160</td>
</tr>
<tr>
<td>Variable</td>
<td>133</td>
<td>111</td>
</tr>
</tbody>
</table>

**6. CONCLUSIONS**

The experiment in this report just serves as an illustration of the ability of the presented technique to compensate for articulator asynchrony. Further experiments in a complete recognition task will show the benefit of the proposed method. The technique is expected to increase the robustness of a recogniser, since it is able to predict infrequent manners of speaking that might not occur in a training material. Much work remains to describe other phoneme boundaries. Our knowledge about their realisation is still incomplete in many ways. Further improvement is dependent on the development of better speech production models. Especially, use of an articulatory model would give a straightforward description of several boundaries, e.g. adjacent consonants. We believe that implementing such a model in the described recognition system would be an important step towards further performance increase.

**7. ACKNOWLEDGEMENT**

This project was supported by the Swedish Board for Technical Development.

**8. REFERENCES**


Including Duration Information in a Threshold-Based Rejector for HMM Speech Recognition

Antonio M. Peinado, Antonio J. Rubio, Juan M. Lopez
Jose C. Segura, and Victoria E. Sanchez
Dpto. de Electrónica y Técn. de Computadores
Facultad de Ciencias, Univ. de Granada
Granada (SPAIN)

ABSTRACT
State duration has been shown as a useful information for recognition. This work tries two ways of including this information in a postprocessing stage, and the effect of incorporating word duration is investigated in each one. In order to diminish the error rate, those utterances that are not clearly recognized can be rejected. Both inclusion ways are tested in a threshold-based rejector. Finally, this rejector is tested with a list of confusing words with those of the vocabulary.

1. INTRODUCTION
In the last years, the HMM techniques reached a very high performance for isolated and connected word recognition and for continuous speech recognition [1]. Applications such as voice dialing of telephones, and automatic credit card entry require a high level of safety. In order to improve the recognition systems accuracy, the three basic stages of such systems (signal analysis, recognition and postprocessing) must be improved. This work is concerned with the study of the postprocessing stage on a speaker-independent isolated word recognition system.

State duration densities can be explicitly incorporated in the HMM algorithms, but the computational cost is quite high. An alternative is to include the duration information in the postprocessing stage, as an additional score to that provided by the HMMs. This solution has been shown as being efficient as the explicit inclusion [1]. In this work, we study two ways of including the state duration in the postprocessing, along with word duration.

For applications that require special safety, a rejection technique can be added in the postprocessing. By means of this technique, those utterances that may yield a misrecognition are rejected. In order to decide a priori which utterance may yield a misrecognition, we propose a detection method that consists on defining a score threshold for each HMM of the vocabulary, and find the best way of including the duration information in this threshold-based rejector.

2. THE HMM-BASED RECOGNITION SYSTEM
The data were sampled at 8.091 KHz, and preemphasized with a preemphasis factor \( \gamma = 0.95 \). Hamming windows were applied to blocks of 256 samples, with an overlapping of 64 samples. Liftered Cepstrum is computed for each frame (with 10 cepstral coefficients and length 12 for the liftering window) and Delta Cepstrum is approximated by linear regression on a ±3 frames environment. Frame energy is normalized to the peak of energy in the word and expressed in the dB scale. Delta Energy is computed from the normalized dB-scaled values of Energy. Finally, an average of all of these parameters is performed every other consecutive frames to compose the feature vectors. The final result is as we had 256-samples frames overlapped 128 samples.

The utterances were coded with a 64-centroid codebook in all the experiments, using the MWDM distance measure [2]. We used one model per word, and, when linear segmentation is used for HMM initialization, 7 states per model.

The vocabulary consists of the ten Spanish digits and the Spanish words (CUERPO, HOMBRE, CODA, MUNECA, MANO, DEDOS), thought for controlling each motor of a Robot.

The database consists of 40 speakers and 3 utterances per speaker and per word (1920 words), and it was recorded under the normal conditions of work rooms, so certain level of noise, such as the computer noise, is included. The conditions of recording (echo and noise conditions) were variable along the time, from the first speaker up to the last speaker. Two subsets of this database were considered for our experiments: a) the first 20 speakers (DB1) for training, and b) the last 20 speakers (DB2) for testing. With this choice, the error rate is not near 0%, because of the variable conditions of the recording, but we can study the variations of the error rates and rejections in our experiments, and simulate a real situation of environment change of the recognition system.

3. INCLUDING DURATION INFORMATION
It is usual to use additional information such as energy and duration, in a postprocessing stage. Since we already use energy information in the feature vectors, we develop our postprocessing only with duration.

State duration and word duration can be included in the postprocessing as two new scores, \( P_{sd} \) and \( P_{wd} \), respectively, where,

\[
P_{sd} = \sum_{i=1}^{N} \log(p_i(d_i))
\]

\[
P_{wd} = \log(p_i(T))
\]

where \( p_i(d_i) \) is the duration distribution of state \( i \), \( N \) is the number of states, \( T \) is the utterance duration, \( P_i(T) \) is a gaussian distribution of word duration.

We consider three ways of calculating the state duration distribution: a) histograms (SD1), b) histograms with normalized duration \( d(T) \) / (T) is the mean word duration (SD2), and c) gaussian distributions with normalized duration (SD3). All of them are tested in the experimental results section. Word duration is easily modeled by a gaussian density, considering the word duration process is a gaussian process (what is basically true). \( \text{SD}_d \) and \( \text{SD}_w \) are incorporated to the word log-score using experimental weights.

4. THRESHOLD-BASED REJECTION
In the postprocessing stage, one possibility to diminish the error rate is to reject those utterances that are not clearly recognized.

Our rejection method consists on defining the a score threshold for each HMM \( \lambda \) of the vocabulary, so when the score \( x \) of a test utterance \( O \) is under the threshold of the recognized HMM, the utterance is rejected. This is possible thanks a temporal normalization of the HMM score \( p(O | \lambda) \) by the word duration \( T \), that extracts the temporal dependence of the HMM score, and, thus, we can compare scores from different utterances (with different durations).

The threshold is \( \bar{x}_\lambda = \overline{x}_{sd} \) and \( \bar{x}_\lambda = \overline{x}_{wd} \), that \( \overline{x}_{sd} \) and \( \overline{x}_{wd} \) are the log-mean and the log-standard deviation, obtained from the training data of a given word. The use of \( \overline{x}_{sd} \) and \( \overline{x}_{wd} \) yields a different threshold for each model \( \lambda \). Moving this threshold (by the factor \( d \) is it possible to get several rejection percentages (RBP2) on the testing database (RDB2=RDB2(0)). In the experimental results section, several experiments are performed to find the best rejection.

5. EXPERIMENTAL RESULTS
As reference, we use a system that provides an error rate of 5.52%, using the HMM score \( P(O | \lambda) \) only. We develop 4 experiments with 4 new types
of score that include duration information. The inclusion of this information is performed in two steps: first, only state duration is included (experiments 1 and 2), and second, state and word durations are included (experiments 3 and 4). These experiments are:

1) Experiment 1: the log-score used for the utterance $O$ in model $\lambda$ is as follows:

$$ x = \log p(O | \lambda) + \alpha_{sd}p_{sd} \tag{3} $$

In this case, the mean log-score per symbol includes the state duration log-score. State duration is included by the experimental weight $\alpha_{sd}$. The optimal error rates for the different $p_i(d_i)$ distributions are: SD1) 4.58% ($\alpha_{sd}=0.7$), SD2) 4.68% ($\alpha_{sd}=0.7$), and SD3) 5.20% ($\alpha_{sd}=1.7$).

2) Experiment 2: the duration information is simply added to the mean symbol score

$$ x = \frac{\log p(O | \lambda) + \alpha_{sd}p_{sd}}{T} \tag{4} $$

The optimal error rates for the different $p_i(d_i)$ distributions are: SD1) 4.79% ($\alpha_{sd}=0.03$), SD2) 4.79% ($\alpha_{sd}=0.03$), and SD3) 5.20% ($\alpha_{sd}=0.03$).

3) Experiment 3: the same as experiment 1, but including word duration information,

$$ x = \frac{\log p(O | \lambda) + \alpha_{sd}p_{sd} + \alpha_{wd}p_{wd}}{T} \tag{5} $$

Word duration information is included as state duration in exp. 1, using an experimental weight $\alpha_{wd}$. An experiment (using SD1, $\alpha_{sd}=0.7$) was developed, obtaining that the error rate is an increasing function of $\alpha_{wd}$.

4) Experiment 4: the same as experiment 2, but including word duration,

$$ x = \frac{\log p(O | \lambda) + \alpha_{sd}p_{sd} + \alpha_{wd}p_{wd}}{T} \tag{6} $$

The optimal error rate is 4.58% for $\alpha_{wd}=0.03$ (using SD1, $\alpha_{sd}=0.03$). These results show that it is better to include the state duration as in experiment 1 than as in experiment 2. The word duration is slightly useful in experiment 4 but not in experiment 3, but, in general, it does not imply any significant improvement. There are no important differences between SD1 and SD2, but SD3 yields the worst results in all the cases. This can be easily understood since state duration is not a gaussian process.

The rejection results of experiments 1 and 2 are depicted in Fig. 1, along with a rejection curve using a non-normalized log-score (all of them with SD1, $\alpha_{sd}=0.7, 0.03$),

$$ x = \log p(O | \lambda) + \alpha_{sd}p_{sd} \tag{7} $$

We can observe that the best rejection is obtained when the duration information is included in the mean symbol log-score (eq. 3), and that the threshold-based rejection works better for low rejections (where the slope curve is higher). Also, the necessity of the temporal normalization for the threshold-based rejection is observed.

![Figure 1: Error rate vs. RDB2 for the log-scores of experiments 1 (+) and 2 (.), and a non-normalized log-score (·).](image_url)

We perform a last trial on the rejection using score (3), SD1 and $\alpha_{sd}=0.7$. It consists on testing the ability of the system on rejecting words that do not belong to the vocabulary. For that, we apply a database (DB3) containing 3 confusible words with every word of the vocabulary (3×16=48 words total). These words are divided in 3 types, according to the number and of phonemes in which the word differs:

- **Type-1**: It differs on one or two consonants.
- **Type-2**: It differs on a vowel.
- **Type-3**: It differs on a vowel plus something else (vowels and/or consonants).

Types 1 and 3 correspond to the closest and farthest words to those of the vocabulary, respectively. Figure 2 shows a plot of the word error rate on DB2 and the mean rejection rate on DB3 (RDB) as function of RDB2(a). We can observe that RDB3 has a good behavior for the same values of RDB2 (the small ones) as the error rate. We can use this graphic to fix a work point of rejection. Table 1 shows, for each type of words on DB3, the percentages of the rejected (R), recognized as correct (C) and recognized as incorrect (U) words (R=3.9, RDB2=5.93). As we could expect, the lowest percentage R corresponds to type 1 words, and the highest one to type 3. Also note that the percentages C and U diminish from type 1 to 3. A important point of this results is that words that do not clearly belong to the vocabulary are rejected quit right.

In figure 3 is depicted a plot of RDB2, RDB3 and RDB3 (rejection on the type 3 subset of DB3) as function of parameter $\alpha$. RDB2(a) has an exponential behavior, while RDB3(a) has a linear one. RDB33(a) keep high in any case.

<table>
<thead>
<tr>
<th>Type</th>
<th>R</th>
<th>C</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type-1</td>
<td>38.4</td>
<td>46.1</td>
<td>15.3</td>
</tr>
<tr>
<td>Type-2</td>
<td>55.5</td>
<td>33.3</td>
<td>11.1</td>
</tr>
<tr>
<td>Type-3</td>
<td>84.6</td>
<td>7.6</td>
<td>7.6</td>
</tr>
</tbody>
</table>

![Table 1: Percentages of R, C and U words for each type of words of DB3.](image_url)

6. SUMMARY

Several HMM log-scores, including temporal normalization and duration information, for utterance evaluation were tested. Among all of them, the best result was obtained using only state duration, including it in the mean log-score per symbol (eq. 3). No significant differences were found between using normalized state duration or not.

A threshold-based rejector (using the proposed log-scores) was used to diminish the error rate in a simple way. It was shown that the temporal normalization of score is basic to perform this rejection. This rejector can be efficiently used to also reject utterances that do not belong to the vocabulary. Logically, the performance of the rejection of a confusing word is better as more different is that word to any of the vocabulary.

REFERENCES


![Figure 2: Error rate (+) and RDB3 (*) vs. RDB2.](image_url)

![Figure 3: RDB2 (+), RDB3 (*) and RDB33 (·) vs. $\alpha$.](image_url)
MODELIZATION OF ALLOPHONES IN A SPEECH RECOGNITION SYSTEM

K. Bartkova & D. Jouvet

Centre National d'Etudes des Télécommunications
LAA/TSS/RCP - Route de Trégastel - 22 300 Lannion - France

ABSTRACT

This paper describes a new approach for modelling allophones in a speech recognition system based on Hidden Markov Models (HMM). This approach allows a detailed modelization of the different acoustical realizations of the sounds with a limited amount of parameters by integrating left and right context dependent transitions as well as acoustical targets. Phonetic knowledge is used in the definition of the structure of the models, and a standard HMM training procedure determines the optimal value of the parameters. The efficiency of the approach is demonstrated both in a multispeaker mode, on a 300 word vocabulary, and in a speaker independent mode on several other databases recorded over telephone lines.

1 INTRODUCTION

The hidden Markov modelling approach is now a widely used technique in automatic speech recognition. Although it allows the optimal parameters of a model for a given training corpus (known words or sentences) to be automatically determined, the structure of the models still remains to be defined manually, and the choice of the "best" basic units is difficult. Basic word units are very suitable for small size vocabularies. But, when the vocabulary size increases, basic sub-word units lead to more compact models. Although phoneme units would be a good theoretical choice, they do not work well in practice, as they do not account for the coarticulation effect due to the context influence. To cope with this problem we had previously developed the pseudo-diphone units [2] which consist of the central part of phonemes, of the transitions between phonemes, and also of some strongly coarticulated sound sequences treated as single units. As an alternative, context dependent units, modelling the acoustical realization of the corresponding sound in a specific left and right context [4], can be used. However, such an approach leads to a large number of models that must be reduced in size, in order to achieve a reliable estimation of the parameters. This can be done a priori, using phonetical knowledge [1], or a posteriori, using some clustering algorithms [3].

In the new approach described in this paper, the Markov models are defined in such a way that they can share as many parameters as possible for modelling the different acoustical realization of any sound. This sharing, based on some a priori phonetical knowledge, allows detailed phonetic distinctions to be introduced in the models, with a limited amount of free parameters that are later determined by an automatic procedure (standard HMM training).

2 MODELLING ALLOPHONES

The new modelization of allophones consists in modelling together, in a single basic unit, all the possible acoustical realizations of a given sound. Each sound is thus represented by a single model, having several entry states and several exit states, and allowing the tying of the probability density functions (pdf's). An example of such a model is represented in figure 1. Each entry or exit state is associated to a specific context, that is, to a class of left or right phonemes having the same acoustical influence on the sound. In this approach, every path from one entry to one exit corresponds to an allophone.

A typical model for the vowels would consist in a shared central portion representing the acoustical "target", and transitions to each entry to the "target", and from the "target" to each exit. However, if necessary, several acoustical targets may be defined and the number of left and right contexts can be increased as much as necessary. Because of the integrated modelization of all the acoustical realizations of any sound, and of the sharing the gaussian pdf's whenever it is possible, a detailed modelization is obtained with a small number of parameters. Thus, they can be reliably determined using a standard HMM training procedure.

2.1 Context Influence

Given that some phonetic environments induce the same coarticulation effects on the adjacent sounds, the entry and exit contexts were defined, for each class of sounds, by grouping together phonemes inducing the same acoustical influence. For instance, consonants sharing the same articulation feature tend to affect the following sound in a similar way. As far as vowels are concerned, the similarity between target positions will closely affect the vowel transition towards the neighbouring sounds.

Vocalic contexts for every allophone: As the tongue position in a semi-vowel production is very similar to that of a vowel production, the vocalic contexts involve vowels as well as semi-vowels. According to point and manner of articulation, 10 relevant vocalic contexts were defined: /i/, /j/, high-front-round, high-mid-round, low, mid-front, mid-back, high-back, front-nasal and back-nasal.

Consonant contexts for vowels, semi-vowel and liquid allophones: Because of the formant transitions they induce on the vowels, as well as on the semi-vowels, the consonants were grouped, according to the place of articulation, in 9 homogenous contexts: labial, labio-dental, dental, alveolar, palato-alveolar, palatal, velar, ltr and /i/. However, the nasal consonants /m/, /n/, /ng/ and /ŋ/ were treated as separate contexts as they may induce a nasization of the following vowel.

Consonant contexts for consonant allophones: The transition between two adjacent consonants is less obvious than between two adjacent vowels. On the other hand, consonants assimilate acoustic features (nasality, voicelessness...) easier than vowels. Thus, the merging of consonant contexts for consonant allophones was slightly different from that used for vowel allophones. 12 relevant contexts were defined according to acoustic features: voiceless plosives, voiced plosives, voiceless fricatives, voiced fricatives, nasals, ltr and /i/.
2.2 Possible "Targets"

The inner part of the models represent the acoustic targets. Thus, in order to take into account the possible assimilation of some of the acoustic context features, several phonological rules were created: representing "standard" pronunciations as well as modified ones, were modelized. The structure of the target models was to allow the modelization of even a rather short duration of the overall sound.

Vowel targets: The following acoustic realization rules were used for the vocalic targets: voiced, partially nasalized, or fully nasalized (not represented on figure 1). The loss of the voiced feature at the beginning or at the end of the sound could occur only in a left or right pause context. In the same way, the nasalized target was accessible only from a left nasal context.

Consonant targets: In the consonantal target modelization, a difference was made between a "normal" non-assimilated target, a pronounced and a partially assimilated target (valid only for voiced consonants) and a nasalized consonant. The partially assimilated target was accessible only in a right pause context and the nasalized target (partially or completely) was valid only after a left nasal context.

Semi-vowel and liquid targets: The structure of the models used for semi-vowels and liquids were very similar to that used for vowels. Nevertheless some specificities separate these two sound classes. One of the main differences consists in the length of target modelizations. As liquids and semi-vowels are sounds realized most of the time with short or even very short duration, and thus are strongly coarticulated with the adjacent sounds, fewer states were attributed to the modelization of their sound targets. Thus 4 "short" targets were used to modelize: a "normal" assimilation without any assimilation effect, a voiced target, a partially assimilated target, and a partially nasalized target.

2.3 Phonological Rules

Besides the coarticulation effects between adjacent sounds treated by the allophone model, a function can handle phonological rules in order to modify the standard phonetic descriptions of the vocabulary words. These rules were used not only to predict specific pronunciation (as in phonology), but rather to tolerate several pronunciations that might occur in a speaker independent mode. Thus, each application of a rule increased the number of possible pronunciations of the words. These explicit phonological rules were the following: 1) each word ending with a consonant and followed by a pause can be pronounced with a neutral schwa like vocable sound after the pause; 2) a voiced fricative preceded by a pause can begin with a very short schwa like vocable sound; 3) a succession of sounds containing a sonorant and the sounds /l/ or /r/ can be realized with an epenthetic neutral schwa like vowel between them (especially in a slow speaking rate); 4) a voiced stop can lose its voiced feature when followed by a voiceless consonant; 5) a voiced stop, followed by a nasal consonant and sharing with it the same point of articulation, can assimilate its nasal feature.

3 EXPERIMENTS

In order to validate this new approach we tested it on several databases recorded over telephone lines. A 500 word vocabulary was used to study the influence of the structure of the vocabulary (number of contexts, usefullness of the targets, etc). This vocabulary was recorded 3 times by 10 speakers, 2 repetitions were used for computing the optimal parameters of the HMM, and the third one was used for testing the recognition performances (in a multispeaker modelization described above has then been applied for speaker independent recognition) to every 16 ms, were used together with their first and second derivatives.

Using the last modelization, described above, and taking into account the temporal derivatives of the acoustic coefficients we finally obtained a 8.44 % error rate on the 500 word vocabulary, which is significantly better than the 11 % obtained with the pseudo-diaphones units on the same database.

3.1 Influence of the structure

In the tests reported in table 1, the acoustical analysis computed every 16 ms a set of 8 coefficients : 6 Mel frequency cepstrum coefficients, the logarithm of the total energy, and its temporal variation. The database used is the 500 word vocabulary (the 500 most frequent French words) recorded by 10 speakers. We report only the error rate on the test set.

The first allophone model used a single simple structure for every sound, involving a single target and 13 exit states. Using a single set of 13 contexts, the same for all the sounds, we achieved a 19.1 % error rate. Introducing the contexts defined in the previous section, and several target models for the sounds, the word error rate decreases to 17.8 %. A further improvement, leading to 16.3 % error rate was obtained by shortening the target model for the liquid and semi-vowel. In the preceding tests, there was no loop allowed on the entry and exit states. By adding these loops, represented on figure 1, longer transition between the adjacent sounds have got a better modelization, and further improvement of the recognition score was obtained with a 14.4 % word error rate.

Table 1 - Error rate on the test set of the 500 word vocabulary for different structures of the allophone models.

<table>
<thead>
<tr>
<th>Structure of the models</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 contexts &amp; 1 target</td>
<td>19.1 %</td>
</tr>
<tr>
<td>More contexts &amp; targets</td>
<td>17.0 %</td>
</tr>
<tr>
<td>Liquids &amp; semi-vowels shorter</td>
<td>16.3 %</td>
</tr>
<tr>
<td>Loops on entry &amp; exit states</td>
<td>14.4 %</td>
</tr>
</tbody>
</table>

3.2 Efficiency of the Approach

In this section, the allophone modelization is compared to the pseudo-diaphones modelization and to the word models. The standard acoustical modelizations computed every 16 ms, were used together with their first and second derivatives.

Using the last modelization, described above, and taking into account the temporal derivatives of the acoustic coefficients we finally obtained a 8.44 % error rate on the 500 word vocabulary, which is significantly better than the 11 % obtained with the pseudo-diaphones units on the same database.

Table 2 - Error rate obtained on several databases (for the test set) with different modelizations: Allophone models (All Pseudo-Diaphones units(PD); and whole word models (Word).

<table>
<thead>
<tr>
<th>Error Rate</th>
<th>All</th>
<th>PD</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digits</td>
<td>0.86%</td>
<td>1.33%</td>
<td>0.69%</td>
</tr>
<tr>
<td>Tregor</td>
<td>1.00%</td>
<td>1.42%</td>
<td>0.86%</td>
</tr>
<tr>
<td>Numbers</td>
<td>4.47%</td>
<td>5.68%</td>
<td>5.68%</td>
</tr>
<tr>
<td>500-Words</td>
<td>8.44%</td>
<td>11.04%</td>
<td>8.44%</td>
</tr>
</tbody>
</table>

The other databases used for the comparisons are: Digits (the 10 digits, recorded by 775 speakers), Tregor (36 French words recorded by 513 speakers), and Numbers (French numbers between 00 and 99 recorded by 740 speakers). Each of them was split in two parts: one half for training, and the other half for testing. For these three databases, the speakers were different in the test and the training set, therefore the reported results (error rate on the test set) corresponds to a speaker-independent model.

As can be seen in the above table, the results achieved by the allophone modelization are significantly better than those obtained with the pseudo-diaphones units. Also, even on small databases, the allophone models, which use less gaussian pdf’s than the word models, lead to performances which are comparable to those obtained with word models.

4 CONCLUSION

The present study described an efficient way of modeling the allophones by representing in an integrated manner all the different possible acoustical realizations of the sounds. Phonetical knowledge was used for the definition of the structure of the models, whereas a standard HMM training procedure determined the optimal values of the model parameters. The application of these allophone models to different databases led to good performances, demonstrating thus the efficiency of this new approach.

REFERENCES

TOWARDS MORE RELIABLE AUTOMATIC RECOGNITION OF THE PHONETIC UNITS

A. Domatas and A. Rudžionis

Technological university, Kaunas, Lithuania

ABSTRACT

This paper is concerned with speaker-independent phoneme recognition in isolated words. We try to evaluate the influence of coarticulation and to create an economic and effective phoneme recognition method which estimates the correlations among features. An adequate evaluation of transitions between phonemes and application of a dichotomization-based (D) classifier permits to decrease the perceivance error for several times as compared with widely used Euclidean (E) classifier.

1. INTRODUCTION

Speaker-independent recognition is so far related to great problems. Comparison of effectiveness of methods widely used [2] does not show essential difference among them. Moreover, the E classifier proves to be equal to other methods. Usually a phoneme is represented by features in its stationary part. Good results are presented in [1] with inclusion of dynamic features when discriminating between nasals. However, these results are obtained on reference set only. Here an approach of automatic estimation of coarticulation and the classifier using an a priori information effectively are proposed.

2. PHONEME RECOGNITION

2.1. Use of coarticulation

Speech signal is represented by consistently following spectral vectors \( S(m) \), where \( m = 1, \ldots, M \) is the number of spectral components, and \( N \) is the number of spectral vector. The instability of every spectral sample is:

\[
\zeta(m) = \frac{1}{N} \sum_{n=1}^{N} E_n(m) \qquad (2)
\]

Spectral vector \( S(m) \) where \( m = \text{argmin} \ e(m) \) corresponds to the stable part of a phoneme and vector of instabilities \( E(h) \), where \( h = \text{argmax} \ e(h) \), corresponds to the transitional part. The logical rules are applied to exclude false extrema. In our experiment, the vector of initial features \( X \) of a phoneme is formed in the three ways:

- from spectrum \( X(S(m))(SP) \); - from spectral vector and vector of instabilities \( X(S(m), E(h))(SPI) \); - from spectrum and consistently following vectors of instabilities \( X(S(m), E(h-2), E(h-1), \ldots, E(h+2))(SPFI) \).

Let \( I, J \in \mathbb{N} \) to be the number of initial features.

2.2. Dichotomization between phonemes

A linear function to discriminate between phonemes \( s \) and \( t \) is used:

\[
g(s) = \sum_{i=1}^{I} w_i x_i^s + Q \left( \text{if } g(s) > 0 \right) = \sum_{i=1}^{J} w_i x_i^t + Q \left( \text{if } g(t) > 0 \right) \qquad (3)
\]

where \( x^{st} \) represents a vector of selected features of unknown test pattern when distinguishing between \( s \) and \( t \) respectively. \( w \) represents a vector of selected reference features of phoneme \( s \). \( W \) is a vector of weights, \( \nu \) is the number of selected features, \( Q \) is a threshold.

For dichotomization of every pair of phonemes the own threshold and sets of features and weights are calculated. During the training we calculate averages \( x_{i1}, x_{i2}, \ldots, i=1, \ldots, \nu \), and correlation matrices \( C^s, C^t \). The Gaussian distribution of features is supposed. Features are ordered according to the decrease of interphoneme distances

\[
\delta_{ij} = \frac{x_i^s - x_i^t}{\sqrt{\sigma^2_{ij}}} \qquad (4)
\]

where \( \sigma_{ij} \) is the variance. Weight \( W_i \) are calculated by using an iterative procedure to minimize the probability of misclassification \( p_f \) \( (J \) is the number of weights already defined). The procedure estimates correlation among features.

Number of selected features \( J \) is computed by:

\[
J = \text{argmax} \ p_f \hspace{1cm} (5)
\]

where \( p_f \) is expected probability of error. \( p_f \) depends on training set size, \( J \) and on \( p_f \), and is defined from the tables in [3].

2.3. Dichotomization-based classifier

Output of an elementary dichotomizer \( O(s) \) is denoted by:

\[
O(s) = \begin{cases} 1 & \text{if } g(s) > 0 \\ 0 & \text{if } g(s) \leq 0 \end{cases} \quad (6)
\]

Respectively, output \( O(t) \) is \( O(t) = 1 - O(s) \). Here two approaches are used to get the final result:

- consistent elimination (D1): class \( t \) is excluded from the list of classes considered if \( O(t) \geq 0 \), and class \( s \) is compared to the next class from the list. The result is the class remained after \( S-1 \) comparisons where \( S \) is the number of classes.

- voting (D2): the result is class \( v \) defined by:

\[
v_{\text{argmax}} O(s) \hspace{1cm} (8)
\]

3. EXPERIMENTS AND RESULTS

3.1. Experimental conditions

- filter bank with 24 or 25 channels; - intervowel spectral features linearly spaced channels; - interval of spectral frames 10 ms; - sampling quantization 8 bits.

3.2. Recognition of stationary vowels

The comparison of D, E and Mahalanobis (M) classifiers was performed to estimate their effectiveness. The speech material consisted of phonemes from words /a/, /e/, /o/, /i/, /l/ spoken by 12 males (4800 patterns). The error rate of reference set recognition (C-examine) and 'leave-one-out' examination (L-examine) is shown in Table 1. Results show that D classifier reduces error rate for more than 4 times in comparison to E one and needs less training than M one for 11 speakers D classifier led to similar error rate for both C and L examines.
In this experiment, D and E classifiers required the similar recognition time.

3.3. Recognition of coarticulated /m/, /n/, /v/, /l/ 

This experiment was performed to investigate the effect of inclusion of dynamic features. The diphones consonant-vowel were selected from words, where vowel was /a, /o, /u, /i/. 11 male speakers took part in this experiment. Reference and test sets consisted of 229 patterns of every coarticulated consonant (2640 patterns in all). The error rate of the test set recognition is shown in Table 2.

Results presented suggest that correct selection of features and use of D classifier provides for recognition error rate less than 42% for all three cases:

- Efficient discrimination of these 4 consonants in context with /a/ is achieved by using stationary part of consonant only;
- In context with /u/ it is necessary to add dynamic features in transition between consonant and vowel;
- Even if the most complicated situation when discriminating among soft consonants, the use of several vectors in transition leads to very low error rate.

3.4. Additional superiority of D classifier

The influence of transmission channel on recognition error rate was examined. One male speaker pronounced 100 patterns of every nasal /m/, /n/ in /a, /o, /u, /i/ context. The hard nasals were pronounced in /a, /o, /u, /i/ context. The soft nasals /m/, /n/ in /a, /o, /u, /i/ context. 100 patterns of every nasal were used for test set by using the same microphone, and by using microphone of another type.

Feature system SP was used. The recognition error of test sets is shown in Table 3. Results show that D classifier is less sensitive when changing the properties of the transmission channel in comparison to E one.

3.5. Automatic labeling of isolated words

The aim of the experiment is the comparison of automatically obtained transcriptions of words to manually formed references, 20 phonemes (50 phonetical subclasses) were used. The alphabet consisted of Lithuanian phonemes except /r/. Stops /p/, /b/, /k/ and /s/, /l/ were united to "unvoiced stop" and "voiced stop" respectively. The labeling process includes two steps. First, the feature system SP is applied. Second, if connection consonant-vowel, nasal-vowel or mixed-vowel is fixed, system SPI is used for more accurate definition of a consonant according to the vowel recognized. 11 male took part in forming reference set, each subclass consisted of 200-1500 patterns. Test set was formed from 50 words spoken twice by 10 males. Average word length was 7.0 phonemes. The correct transcription of a word was fixed if it adequately coincided with the transcription of its reference. The test led to 32% correct transcriptions of words for D classifier and to 6.2% for E one correspondingly.

4. CONCLUSIONS

We have presented two methods to improve phoneme recognition. Inclusion of dynamic features into representation of phonemes provides for significant decrease of recognition error rate. Dichotomization-based classifier offers the following advantages:

- Inclusion of essential features only for dichotomization between phonemes;
- Selection of feature set guaranteeing minimum probability of dichotomization error;
- Material influence of transmission channel because of effective application of correlations among features;
- Less training set necessary to form representative features in comparison to Mahalanobis one;
- Less recognition time required in comparison to Mahalanobis one;
- Lesser error rate for several times in comparison to Euclidean one.

5. REFERENCES


Table 1

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Number of features</th>
<th>Examike</th>
<th>NS=1</th>
<th>NS=4</th>
<th>NS=11(12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>/3=8</td>
<td>C</td>
<td>5.3</td>
<td>8.2</td>
<td>12.2</td>
</tr>
<tr>
<td>M</td>
<td>/3=8</td>
<td>C</td>
<td>0.6</td>
<td>0.7</td>
<td>1.8</td>
</tr>
<tr>
<td>D1</td>
<td>/3=4 (/3=8)</td>
<td>C</td>
<td>14.3</td>
<td>10.6</td>
<td>6.6</td>
</tr>
</tbody>
</table>

NS is the number of speakers used for reference forming.

Table 2

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Method of phoneme representation</th>
<th>Number of features</th>
<th>Vowel of diphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>SP</td>
<td>/3=24</td>
<td>/3=16</td>
</tr>
<tr>
<td>M</td>
<td>/3=8</td>
<td>3.0</td>
<td>5.6</td>
</tr>
<tr>
<td>D1</td>
<td>/3=12 (/3=24)</td>
<td>2.8</td>
<td>8.3</td>
</tr>
<tr>
<td>D2</td>
<td>/3=12 (/3=48)</td>
<td>1.3</td>
<td>3.8</td>
</tr>
<tr>
<td>D3</td>
<td>/3=12 (/3=14)</td>
<td>-</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Table 3

<table>
<thead>
<tr>
<th>Classifier</th>
<th>The former microphone</th>
<th>Another type microphone</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>4.0</td>
<td>13.9</td>
</tr>
<tr>
<td>D1</td>
<td>1.5</td>
<td>2.6</td>
</tr>
</tbody>
</table>
THE SYLK PROJECT: SYLLABLE STRUCTURES AS A BASIS FOR EVIDENTIAL REASONING WITH PHONETIC KNOWLEDGE

P.J. Roach*, D. Miller*, P.D. Green* and A.J. Simons*

*University of Leeds
Leeds LS2 9JT, U.K.

University of Sheffield
Sheffield S10 2TN, U.K.

ABSTRACT

This paper reports on work being done on the SYLK project, funded by the UK IEATP programme (project no. 1067); this is aimed at developing a syllable-based speech recognition system combining statistical and knowledge-based approaches to sub-word unit recognition, suitable as a front-end for large-vocabulary, speaker-independent applications. Hidden Markov Models are used to construct initial hypotheses for the knowledge-based component; encouraging results in recognising different sub-word units are presented.

1. INTRODUCTION

The sub-word unit on which SYLK is based is the syllable; the acronym stands for 'Statistical Syllabic Knowledge'. An overview of the whole project is given in [5]. The arguments in favour of syllable-based recognition are well-known ([1]); the principal reason for choosing the syllable (and this is true to a lesser extent of the demisyllable and triphone units) is that much of the allophonic variation found in phonemes can be explained in terms of the syllabic position in which they occur. An example is the difference between voiced /t/ and its voiceless allophone [t] found after /p t k/ in words such as 'pray', 'tray', 'cray': a phoneme-based recogniser trained to recognise /t/ would need to be trained to recognise voiced and voiceless allophones separately, whereas a system trained to recognise syllable onsets of the form voiceless stop, /r/ would not need to be given variants: a voiceless /r/ is simply a normal property of syllables beginning in this way.

The motivation for the combined statistical and knowledge-based approach is that recognition by statistical model alone seems to work very well for the majority of straightforward instances of the units being recognised, though it is critically dependent on the initial training data; knowledge-based systems, on the other hand, have the ability to make use of multiple sources of knowledge to refine hypotheses at more and more detailed levels, but risk becoming fatally derailed if the initial hypotheses with which they start are incorrect. The ideal strategy therefore seems to us to be one which embodies a statistical component for making initial hypotheses, and a knowledge component for hypothesis refinement. In this approach, it is more important for the initial hypothesis not to be wrong than for it to be exactly right in full detail.

This paper is chiefly concerned with the initial, statistically-based part of the system, this being the one which has been most fully developed at the present time. In the full SYLK system, the lattice of SYLK symbols provided from the first pass is used to instantiate (independently) hypotheses about the structure of each syllable in the utterance, centred on its peak. Allowed syllable structures, and their interrelationships, are made explicit by an object-oriented Syllable Model; further processing is based around the application of 'refinement tests' to the syllable structure hypotheses ([2]).

2. CHOICE OF UNIT FOR INITIAL HYPOTHESIS CONSTRUCTION

For large-vocabulary speech recognition, the most convenient form of output from the front-end is a phoneme lattice allowing subsequent lexical access from dictionary entries coded in terms of phonemes (though other lexical access techniques can be used). For the reasons explained above, however, we prefer not to work with phonemes as our recognition unit within the front-end: instead we envisage that the final stage in our front-end processing will be to recover a phonemic transcription from the syllable-based, allophonic explanation which SYLK will produce. Although our explanation unit is the syllable, there is no reason why we should not build initial hypotheses on the basis of phoneme-sized units if they can be reliably recognised. We may, for example, segment and label the speech signal in terms of acoustic phonetic units, where all major allophones of the phonemes are identified in a context-free manner. Alternatively, we may choose to identify phonetic segments that are members of a much smaller set: such broad phonetic categories (often based on manner of articulation, comprising categories like plosive, fricative, vowel, nasal) are likely to give more robust recognition (see [8],[10]). Another possibility is to attempt to recognise units above the level of the phonetic or phonemic segment. It is generally agreed that the number of syllables used in English exceeds 10,000, and to develop statistical models of all of these would not be computationally practical; consequently a unit smaller than the syllable may be best. Triphone modelling is used, for example, ARMADA ([11]); another unit which has its supporters is the demisyllable ([4],[12]).

For our purposes, bearing in mind that we are working towards decoding speech into fully-specified syllables at a later stage in the process, we prefer to make use of smaller units than demisyllables, but units which are explicitly tied to syllabic structure (which diphones and triphones are not). It is usual to view the syllable as composed of an optional ONSET, an obligatory PEAK (normally the vowel) and an optional CODA, each of which can be treated as independently recognisable objects ([1]). We believe there to be approximately 60 possible Onsets in English and about 120 Codas, while the number of Peaks is in the region of 20. Strangely, there appears to be no phonological term for referring in a generic way to Onsets, Peaks and Codas, and we are reduced to calling them Syllable Constituents. Although these units are potentially useful, we have chosen to work with units of the same size as Syllable Constituents but less fully specified. For example, we believe it to be unrealistic to expect a straightforward statistical recogniser to achieve speaker-independent, context-free discrimination of [sp], [st], [sk], [hpi, fki], etc; but we do think it feasible to aim to recognise the class of [sp], [st], [sk], etc. If we bring together on acoustic grounds all highly-confusable Onsets and, separately, Codas into broader units, we reduce the set of Onsets to 30 and of Codas to 60. Again, no name exists for such units, but we have come to refer to them as SYLK units ([9]).

3. EXPERIMENTS IN STATISTICAL
RECOGNITION OF SUB-WORD UNITS

We have been careful throughout this work to make use of widely-available and
widely-used speech data and performance testing techniques so that our results
should be comparable with research done elsewhere. Our original intention was to
make use of a British English database as envisaged in the SCRIBE project, but
delays in the production of this has obliged us to use instead the TIMIT
corpus of American English. Since the
total amount of data recorded on the
current TIMIT CD-ROM disk is very
large (4300 sentences spoken by 420
speakers), we have made use of a subset
for training and testing purposes, based on the 1030 sentences collected from
Dialect Regions 1 and 7, we discarded "duplicate" (SA) sentences and ones with
obvious transcription errors. Two
sentences from each speaker were kept as
test data, the remainder being used as
training data. Female and male voices are
being studied separately at present, and
full results for the female voices are not
yet available.

We have conducted a series of
experiments in recognizing sub-word units.
Two different units were chosen, one a
phone-sized unit based on the segments
labeled in the TIMIT corpus, and the
other the SYLKunit as described above.
For the former, we trained models on
every phonemic category. However, in its
most detailed form, the TIMIT
transcription describes the "slight portion" of /p/, /t/ and /k/ which is
clearly not practical; by ignoring errors
within such categories we effectively aimed
at recognition at a level known as
"reduced TIMIT" ([7]), roughly
comparable in detail with phonemic representation. We have also tried "broad
class" recognition of the same-sized unit.

Since no corpus annotated with
SYLKsymbols was available, we had to
produce our own. While some material
in British English has been specially recorded and
transcribed to give a full coverage to
all possible Onsets and all possible Codas,
our current use of American English and
our need for large quantities of training
data made it necessary to carry out an
automatic re-coding of the TIMIT data
into SYLKsymbols. This was done,
making it possible to train HMM's for
recognition of two different types of unit
on the same recorded material. Since non-
peak SYLKunits are characterized as
Onset or Coda, the re-coding required
decisions about syllable boundaries; as is
usual, such decision were based on the
Maximal Onsets principle according to
which all inter-vocalic consonants are
assigned to the Onset of the following
syllable if this does not violate phonotactic
regularity.

It is essential to have a reliable and
meaningful technique for scoring the
recognition success rate. For work using
TIMIT it has been usual to use the
scoring technique developed at NIST for
work on TIMIT, and we originally used
this. We have recently adopted as our
standard HMM software resource the
HTK package developed at Cambridge
University Electrical Engineering
Department, and this contains a scoring
technique that is similar to the NIST test.
All our results given below were calculated
by HTK scoring; we observe the standard
scoring distinction between correct and
accurate (where in the latter case, insertions cause a reduction of the score).

4. RESULTS

4.1 Recognition Scores

At the time of writing, the best scores we
have achieved on the TIMIT test data are
shown in Table 1 (data from male
speakers only):

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Accurate</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>56.0%</td>
<td>51.6%</td>
</tr>
<tr>
<td>SYLK</td>
<td>67.9%</td>
<td>53.5%</td>
</tr>
<tr>
<td>SYMBOLS</td>
<td>60.8%</td>
<td>57.2%</td>
</tr>
</tbody>
</table>

It is important to compare these with
results from elsewhere; the closest
correlation we have been able to find is
the context-independent phone recognition
on TIMIT data reported in [7] using male
and female data, they reported 54%
Correct and 33.2% Accurate. Glottal stops
were ignored in their study, whereas we
have treated these in the ways of the phones to
be recognized.

4.2 Comparative Evaluation: Phonetic Segments vs. SYLK units

There remains an unsolved problem in
interpreting these results: the two units
studied are in some ways radically
different from each other, and are not
easily comparable. While excellent
methods exist to compare two different
attempts at recognition of a particular set
of units in an utterance (e.g. [3]), what we
have here is scores for units of different
categories and containing different amounts of
information. We need to know which of
the two units brings us principle closest to
successful word recognition. One way of
doing this is that we are currently
investigating is first to discover which representation gives less uncertainty in
word identification, using an approach
based on [6]. We are using an on-line
pronouncing dictionary of approximately
70,000 words and automatically re-coding
the entries in SYLKsymbol and in
TIMIT phonemic symbols. Each word, in
both representations, will then be checked
against all the others to see how
many other dictionary entries have
identical coding, and the representation
showing the smallest number of
confusions will be shown to be the most
favourable for word recognition. It should
be remembered, however, that much might
be gained by supplying the knowledge-
base component of SYLK with both
representations as part of a classification system.

5. REFERENCES

the same data". Memorianda 41/6, RSRE Malvern.
phonetic information", Proc.ICSLP
RECOVERING TUBE KINEMATICS USING TIME-VARYING ACOUSTIC INFORMATION

R. S. McGowan
Haskins Laboratories, New Haven, Connecticut

ABSTRACT
Formant frequency trajectories are used to optimally fit the kinematics of a modified twin-tube model. An entire articular trajectory is fit in a single optimization, because an articular trajectory is modeled as a parameterized function of time.

1. INTRODUCTION
The inverse mapping between acoustics and articulation has received considerable attention in the last twenty-five years. Most of the work has been on mapping static spectral variables onto static vocal tract shapes, with resulting ambiguity in the mapping. Ambiguities were noted in the work of Atal, Chang, Mathews, and Jukey [1] where the articulatory positions of the vocal tract model were varied to fit formant frequency data; in the work of Flanagan, Ishizaka, and Shipley [3] using an optimization procedure based on spectral information and cepstral matching to find vocal tract area functions, as well as vocal tract pressure and laryngeal parameters; and in the work of Levinson and Schmidt [5] using a gradient search optimization to relate articulatory and acoustic envelopes.

Two ways of overcoming inverse mapping ambiguities suggest themselves: either decrease the number of articulatory degrees of freedom, or increase the amount of acoustic data. One procedure to decrease the number of articulatory degrees of freedom took account of the continuity of vocal tract tube shapes in short time intervals [6,4,8]. This seemed to help relieve ambiguity, but the optimizations were performed at each time sample, making the inclusion of the continuity constraint inefficient. In the method examined here, the kinematics of the articulators were parameterized as functions of time, and the optimization was performed over time spans corresponding to a single parameterization, thus the continuity constraints were automatically incorporated. Because the time spans were longer than a single time sample, there was a span of acoustic data that was used in the optimization, thus the number of degrees of freedom in the data was also increased.

2. METHOD
The acoustic data consisted of up to three formant frequency trajectories that were generated using a modified twin-tube model [2]. In the modification considered here, a third tube, a constriction tube, was placed between front and rear tubes of the twin-tube model (Fig. 1). There were five articulatory variables: front tube area, constriction tube area, rear tube area, rear tube length, and constriction tube length. The front tube length was determined by the restriction that the total tube length be 17 cm. The constriction area was parameterized as an exponential function of time. The maximum area of the constriction was assumed to be the average of the front and rear tube areas, and the minimum was zero, corresponding to complete constriction. As a result there were five articulatory kinematic parameters: the four constant articulatory variables, and the exponential growth factor for the change in constriction area (Fig. 2).

The modified twin-tube model was used for both the synthesis of formant frequency data and as a model vocal tract for articulatory kinematic parameter recovery. The relationship between the acoustic variables and the articulated variables was given by the model function. This function was written as an implicit relation between the formant (frequency) functions and the articulatory variables. Thus, if the constriction area was given a trajectory, either opening or closing, it is possible to compute the corresponding formant trajectories using numerical root-solving techniques.

![Figure 1: Modified Twin Tube](image)

![Figure 2: Articulatory Kinematic Parameters](image)

Preliminary work has been done on recovering articulatory kinematic parameters from synthesized formant frequency trajectories using the modified twin-tube model using a least-squares criterion. The iterative least squares was performed using the simplex method [7]. The simplex method was a conservative choice because it did not require numerical computation of a generalized inverse, as, say, the Levenberg-Marquardt algorithm did, thus reducing the possibility of numerical instability in this initial study. However, the simplex method was very slow and could be replaced with more sophisticated optimization algorithms. When the experimenter executed the program written for inverse mapping he was asked to specify the constriction length and was given the option of specifying either the front or rear tube areas. If neither of these was specified, then the optimization was performed to find four parameters: the front and rear tube areas, the rear tube length, and the exponential time constant. If one of the areas was specified, then the optimization was performed on three parameters, and if both areas were specified, then two parameters entered into the optimization: front tube area and the exponential time constant. Because the optimization procedure was an iterative procedure that could be trapped in local minima, the simplex method was run from several initial starting places in the articulatory kinematic parameter space. The search from any of these initial starting places would terminate if the cost function was less than a given tolerance, if there was little relative change in the value of the cost function from one step to another, or if a maximum number of iterations was attained.

The ideal cost function was the sum of squares of the differences over time in each formant frequency between those given by the data and the values that would be produced by the modified twin-tube model given the articulatory kinematic parameters. To have found the value of this cost function, corresponding to a given set of articulatory kinematic trajectories, would have had to be found. This would have involved applying root-solving techniques to the model function many times (40 times for each formant at a rate of 200 Hz for 200 msec). Accordingly, the sum of the squares of the model function evaluated at each data formant frequency was used as an alternative cost function. This appeared reasonable because it is a necessary condition that this function, being an implicit relation between formant frequency and articulatory variables, be identically zero, if the original cost function is zero.

3. RESULTS & CONCLUSION
In the modified twin-tube model, the feasibility of fitting rear tube length and exponential time constant was tested using the first formant frequency trajectory only, as well as with three formant trajectories. The feasibility of fitting four parameters, the rear tube area, front tube area, rear tube length, and exponential time constant using one and three formant frequency trajectories was also tested. As one would expect, the method did better in fitting two parameters than it did in fitting four
parameters. A counter-intuitive result is that the method seemed to have worked better with one formant (e.g., Fig. 3) than it did with three (not shown), or with less information than more. (The program was completely unsuccessful at fitting four parameters given three formant frequencies.)

It was felt that something of the original cost function involving the squares of the differences between formant frequency data and those which would be produced with a given set of articulatory kinematic parameters had to be preserved to get better results. Instead of root-solving for all the formant frequency values corresponding to a given set of articulatory kinematic parameters, root-solving was performed only at the beginning, middle, and end of a trajectory for each iteration of the least-squares procedure. (For example, there were nine root solves for three formants.) The sum of squares of the differences between these frequency values and their corresponding data points were added to the sum of squares of the model function evaluated at all the data points to form a hybrid cost function. This seemed to have alleviated the counter-intuitive result of doing more poorly with three formants (Fig. 5) than with one (Fig. 4). Also, it was possible to fit the four parameters using three formant trajectories (Fig. 5).

The problem with using just the sums of squares of the model function in the cost function was that local minima appeared that were not close to the articulatory kinematic parameters that produced the data. By adding some explicit information to the cost function these superfluous minima no longer hindered the algorithm.

Work supported by NIH Grant HD-01996 to Haskins Laboratories.

4. REFERENCES

---

![Figure 3: One resonance frequency trajectory, implicit function minimization](image1)

![Figure 4: One resonance frequency trajectory, implicit function & frequency difference minimization](image2)

![Figure 5: Three resonance frequency trajectories, implicit function & frequency difference minimization](image3)
PHONEME-LIKE MODEL OF SPEECH SIGNAL

S. Noreika and A. Rudziénik

Technological university, Kaunas, Lithuania

ABSTRACT

Phoneme-like representation of speech signal for single speaker isolated word recognition is discussed. Speech signal is divided into transitions and stationary parts by estimating a spectral instability function. The number of these parts is close to the number of phonemes. The comparison of reference and test patterns is based on processing of the similarity matrix. Well known dynamic time warping technique as well as our original technique are used. The recognition error rate is 0.9% (vocabulary size - 200 words, memory requirement - less than 40 bits per word).

INTRODUCTION

The problems of speaker-independent speech recognition are well known. The employment of a priori information at the phonetic level is supposed to be the effective mean for speech recognition. New methods for phonemes detection [1] ensure high accuracy and error rate several times less as compared to other methods widely used in speech recognition technology. Nevertheless, the algorithm of the phonetic recognition of words is not clear enough. Our purpose is to discuss a phoneme-like model for single speaker speech and to evaluate the main parameters responsible for recognition results.

According to our model transitions and stationary parts of speech signal are detected. The number of these parts is close to the number of phonemes. It is achieved by estimating a spectral instability function. Extremal values of this function after filtering and thresholding procedures correspond to phonemes. On the stage of comparison of reference and test patterns we tried to find estimations, which were more efficient as compared to, e.g., dynamic time warping (DTW) technique estimation. To evaluate coarticulation phenomena, multiple reference patterns per phonetic unit were used. Model testing resulted in 0.9% error rate of speaker-dependent recognition of 200 words. The phonetic transcription of a word was its reference and less than 40 bits of memory was required for its storage.

2. PHONEME-LIKE REPRESENTATION

2.1. Speech signal segmentation

The task is to divide speech signal into transitions and stationary parts. It is desirable that the number of these parts was close to the number of phonemes. In other words, a phoneme-like model is required. The method of selecting transition and stationary frames is based on the estimation of a spectral instability function. Let \( S_{n} \) represents a set of logarithmic spectral vectors of speech signal, where \( n \) denotes the discrete time instant and \( L \) denotes the number of a spectral component. The spectral instability function may be defined as follows:

\[
\beta_{L} = \sum_{n=1}^{N} \left| \frac{S_{n+1} - S_{n}}{S_{n}} \right|
\]

where \( |a_{n}| \) is the interval of spectral instability estimation, \( \eta_{n} = 2 \).

The main segmentation function is

\[
\beta_{L} = \frac{1}{L} \sum_{i=1}^{L} |\beta_{L}^{i}|.
\]

The maxima of this function are related to transitions while the minima are related to stationary parts of signal. To eliminate extremal values of the segmentation function which are related to local fluctuation of spectral parameters, filtering and thresholding procedures of phonemes are adapted. The number of consecutive pairs "transition-stationary part" \( \gamma \) characterizes the extent of compression. Ideally, \( \gamma \) should be equal to the number of phonemes in a word.

2.2. Representation of feature vectors

Spectral instability coefficients and spectral parameters are used for description of transitions and of stationary parts accordingly. Both reference and test patterns may be represented as follows:

- feature vector consists of a set of successive frames, corresponding to transitions and stationary parts (the first phoneme-like model (PLM1));
- transition and stationary part following are treated as a single component of a feature vector (the second phoneme-like model (PLM2));
- vector quantization (VQ) may be applied to PLM1 or PLM2.

2.3. Phoneme verification model

The modification of the phoneme-like model is possible. We call it the phoneme verification model (PVM). The main distinguishing features of the PVM are: (1) the phonetic transcription of a word is its reference, (2) a database characterizes each element of the phonetic alphabet used, and (3) transitions are not used.

A database contains a set of spectral parameters of each phonetic element. Several versions may be used for representation of each phoneme. The clustering technique is very suitable for this purpose.

3. Comparison of test and reference patterns

One of two matrices can be used for the comparison of reference and test patterns: (1) a matrix of local distances and (2) a matrix of local similarities. We consider a similarity matrix is preferable to a distance one. A similarity matrix is supposed to have more information than DTW algorithm uses. Let \( d_{ij} \) is an element of a distance matrix, then an element of a similarity one is defined as:

\[
\delta_{ij} = \frac{1}{1 + d_{ij}}.
\]
The measure of similarity was first used in "leave-one-out" tests. The recognition algorithms consisted of analog/digital converters. All channels were sampled every 10 ms by 8-bit analog/digital converters. The vocabulary consisted of 79 graphic symbols, i.e., on the average, a word consisted of 7.94 letters. The extent of compression of various segmentation algorithms was evaluated on the base of this figure. In the recognition experiments the reference and test patterns were chosen according to the "leave-one-out" procedure, obtaining a total of 9000 tests. In some experiments only part of these tests was used.

4.2. PLM1 and PLM2 testing Several variants of recognition were investigated. The first two methods were the usual DTW methods on the basis of a local distance matrix (V1) and of a local similarity matrix (V2). The third variant V3 differed from V2 by the normalization of the integral similarity measure according to the average duration of the reference and test patterns. The variants V4 and V5 are similar to the variants V2 and V3, but the former was the only three-side-by-side diagonals of the similarity matrix having the largest similarity. The logical processing of elements belonging to these diagonals is the essence of the sixth variant V6. Finally, the seventh variant V7 was the modification of the variant V6, including the segmentation errors correction. Feature vectors for the variant V1 were represented according to PLM1, while the other variants used representation according to PLM2. The results are presented in Table 1, where $N_t$ is the number of test patterns and $E$ is the recognition error rate.

Our model ensures high extent of compression and the number of detected phonemes $Y$ is close to the average number of letters (7.94). Generally, the recognition error rate is inversely related to the extent of compression. The optimization of the integral similarity measure and the employment of diagonals reduces the recognition error rate. The variants V6 and V7 give the best results and these results are achieved without using DTW algorithm.

4.3. Vector quantization A 128-element codebook was generated for PLM1 (memory requirement was about 100 bits/repetition) and for PLM2 (memory requirement was about 50 bits/repetition). The recognition results are shown in Table 2.

Naturaly, VQ reduces the recognition accuracy, nevertheless, the results are high enough on condition that such an extent of compression is used.

4.4. PVM testing To test this model, a 200-word vocabulary was used. As mentioned above, the phonetic transcription of a word was its reference. In the recognition experiment vocabulary was read 7 times, i.e., the total number of tests was 1400 words. The database was formed by clustering speech material containing 50-100 repetitions of each phones. Some phonetic units were considered as one phones, e.g., /p,t,k/ or /b,d,g/, so only 16 phonetic units were used. Hence it follows that memory requirement was only 4n bits per reference, where $m$ is the number of phonemes in a word. The recognition was carried out according to variant V6, except that only two diagonals were used. The results are shown in Table 3. The model gives the promising results. They are conditioned mainly by the use of a priori information about phonemes and by the proper processing of the similarity matrix. Note the main attractive features of this model: (1) practically extrem compression of speech is achieved, (2) once the database has been formed it may be used with any vocabulary, (3) the amount of similarity calculations does not depend on vocabulary size and (4) vocabulary can be changed easily.

CONCLUSION The models used here ensure high extent of compression of speech signal without degradation of useful information. Recognition of 200 words showed that recognition error rate was 0.9% and memory requirement was less than 40 bits per reference. In the future these models are supposed to be used for speaker-independent speech recognition.

REFERENCE [1] DOMATAS, A. and RUDZIONIS, A. "Towards more reliable automatic recognition of the speech material. Note the main attractive features of

Table 1: Comparison among various variants of recognition

<table>
<thead>
<tr>
<th>Variant</th>
<th>Nt</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>9000</td>
<td>2700-3300</td>
<td>9000</td>
<td>9000</td>
<td>9000</td>
<td>9000</td>
<td>9000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.1-7.4</td>
<td>7.2</td>
<td>7.2</td>
<td>7.2</td>
<td>7.2</td>
<td>7.2</td>
<td>7.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.4-6.0</td>
<td>6.0</td>
<td>3.4</td>
<td>4.3</td>
<td>2.3</td>
<td>1.6</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 2: PLM1 and PLM2 testing results with VQ

<table>
<thead>
<tr>
<th>Nt</th>
<th>Without memory requirement, bits</th>
<th>with memory requirement, bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>900</td>
<td>0.5</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>VQ</td>
<td>VQ</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Représentation de connaissances indépendantes du locuteur pour la reconnaissance de mots acoustiquement proches

H. Méloni, P. Gilles

Laboratoire d’Informatique - Faculté des Sciences
33 rue Louis Pasteur - 84000 Avignon - France

**RESUME**

Nous proposons une méthodologie pour la discrimination descendante entre des mots phonétiquement proches d’une cohorte. Les connaissances utilisées ne dépendant que de quelques caractéristiques très limitées du locuteur (position des formants pour les voyelles) et décrivent les traces acoustiques de phénomènes articulatoires dans un contexte connu. Ces techniques sont appliquées à l’identification des occlusives sourdes dans des logatomes constitués des consonnes /pl/, /bl/ et /kl/ suivies d’une des voyelles du français.

1. PRESENTATION DU PROBLEME

Le Décodage Acoustico-Phonétique de la parole est rendu difficile notamment à cause des variations inter-locuteurs et des effets de la coarticulation des phonèmes. Le premier type de variabilité, de nature statique (cibles différentes), peut être traité partiellement de manière ascendante par l’utilisation de quelques caractéristiques d’un locuteur (modèle spectral des parties stables des unités phonétiques). L’acquisition, la mémorisation et le traitement de ces connaissances sont aisément effectués et permettent de mettre en œuvre une première phase efficace du DAP [2], [4], [5]. Les résultats d’un tel processus sont constitués par un treillis de phonèmes valant toutes les hypothèses vraisemblables d’occurrence d’une unité. Ces éléments déterminent des ensembles de mots qui sont susceptibles de coïncider de manière optimale - au sens de critères de proximité et de densité de recouvrement - avec une zone du treillis. Les mots proposés dans la phase ascendantante sont acoustiquement proches et les scores de reconnaissance qui leurs sont associés ont été calculés au moyen de distances par rapport à des références idéales non altérées dans le contexte. Il convient donc, dans une phase descendante du processus de décodage, de classifier plus précisément ces hypothèses. Les phénomènes de coarticulation ont pour conséquence la modification des cibles phonétiques et apparaissent sur l’évolution temporelle des paramètres acoustiques et phonétiques (formants par exemple). La phase descendante du DAP consiste à localiser et évaluer les traces acoustiques de phénomènes articulatoires distincts sur les zones appropriées du signal. Cette opération est effectuée en utilisant les connaissances disponibles sur le contexte phonétique. Les travaux présentés ici décrivent la méthodologie utilisée et les résultats obtenus pour la discrimination des occlusives sourdes dans le cas où les mots sont des logatomes constitués d’une consonne suivie de l’une quelconque des voyelles du français. Nous examinerons plus particulièrement le processus d’identification du lieu d’articulation.

2. METHODOLOGIE

L’identification du lieu d’articulation des occlusives sourdes peut être effectuée au moyen d’informations diverses (spectrales et temporelles) qui apparaissent sur l’explosion et dans la transition vers la voyelle adjacente [2], [3], [7]. Nous envisagerons que les traces acoustiques détectées sur les paramètres spectraux.

2.1. Paramétrisation du signal

Le signal de parole est numérisé sur 16 bits à une fréquence de 12.8 kHz puis préaccentué et caractérisé chaque 10 ms par son énergie globale, la densité des passages par zéro et les énergies spectrales dans 24 canaux répartis suivant une échelle de Mel. Les spectres sont obtenus par prédicteur linéaire et cette représentation est suffisamment efficace pour représenter la plupart des connaissances. Il est cependant parfois indispensable de disposer de paramètres plus précis, notamment pour suivre les trajectoires formantiques. Dans ce cas, nous disposons d’une caractérisation plus fine des spectres LPC (figure 1). Un ensemble d’outils permet de définir et de calculer dynamiquement de nombreux paramètres auxiliaires obtenus par combinaisons des attributs initiaux [5]. Les informations les plus utilisées mesurent et comparent les densités d’énergie dans certaines bandes spectrales. L’évolution temporelle de ces paramètres est modélisée au moyen de formes élémentaires.

Les calculs de $E(p,v)$ sont effectués à partir des valeurs de la table 1. L’énergie correspond à celle du canal d’onde où l’amplitude est maximale dans la zone spectrale indiquée par la règle associée à la situation donnée.

2.3. Identification sur la transition

Pour modéliser les informations relatives à l’évolution spectrale de l’énergie autour des formants, il est nécessaire d’utiliser une représentation des spectres au moyen de 128 valeurs (figure 1). Toutefois, la caractérisation au moyen des 24 canaux permet de mesurer les évolutions temporelles des formants dans le cas où les valeurs des formants de chacune des voyelles pour un locuteur donné.

Pour les occlusives /pl/, /bl/ et /kl/ des règles définissent et calculent les attributs caractérisant l’énergie spectrale de la composante principale de bruit d’explosion en fonction de la position des formants de la voyelle adjacente. Si nous notons $E(p,v)$ la densité d’énergie dans la zone désignée pour la consonne $p$ dans le contexte de la voyelle $v$, nous pouvons calculer la fonction

$$f(p,v) = 2\times E(p1,v) - E(p2,v) - E(p3,v)$$

qui définit la valuation de l’hypothèse correspondant à la consonne $p$. La valeur de la fonction est d’autant plus grande que la position spectrale du bruit coïncide avec celle définie pour cette situation.

Table 1 - Position du bruit d’explosion pour les occlusives sourdes en fonction de la position des formants de la voyelle.

<table>
<thead>
<tr>
<th>/p/</th>
<th>/b/</th>
<th>/k/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/pl/</td>
<td>/F2+1</td>
<td>/F3+2</td>
</tr>
<tr>
<td>/bl/</td>
<td>/F2+1</td>
<td>/F3+2</td>
</tr>
<tr>
<td>/kl/</td>
<td>/F2+1</td>
<td>/F3+2</td>
</tr>
</tbody>
</table>

Figure 1 - La représentation spectrale au moyen de 128 canaux est nécessaire pour permettre le suivi précis de formants.

2.2. Identification sur l’explosion

Dans la phase de DAP ascendant, la position de l’explosion a été repérée au moyen de paramètres calculés en fonction du phonème. Nous disposons par ailleurs...
pôles significatifs sont suffisamment séparés. La direction de la transition des formants est évaluée sur la portion de la voyelle située entre le début d’apparition des pics spectraux et la trame de plus grande stabilité. Le calcul des valeurs de la pente du formant (repéré par le canal /i/ au maximum de stabilité) est effectué à partir de l’évolution de l’énergie dans les canaux adjacents (canaux /i-1 et /i+1). La différence de densité d’énergie entre la zone stable et le début d’apparition des formants dans les canaux /i-1 et /i+1 constitue le paramètre essentiel permettant d’apprécier le sens de l’évolution d’un formant au contact de la consonne (figure 2).

Figure 2 : Les canaux /i-1 et /i+1 sont utilisés pour mesurer l’évolution temporelle de l’énergie autour du formant (canal i).

Les informations concernant les transitions sont utilisées pour compléter celles qui sont évaluées sur l’explosion. Nous avons limité ces connaissances aux seules situations qui sont pertinentes pour de nombreux locuteurs et qui peuvent être traitées à partir de la représentation paramétrique sur 24 canaux. Les formes des transitions de référence utilisées sont données dans la table 2. Il s’agit d’une tendance générale plus ou moins marquée suivant le contexte et le locuteur. Ces indices acoustiques traduisent l’influence du lieu articulatoire de la consonne sur la cible de la voyelle.

Table 2 - Formes des transitions des formants /F2 et /F3 pour les voyelles précédées des oscillations soudées. Seules les formes utilisées dans notre système pour l’identification du lieu articulatoire sont présentées dans cette table.

3. RÉSULTATS
Les règles ont été testées sur un corpus étiqueté automatiquement (étape ascension du DAP) de plusieurs centaines de phonotomes prononcés par 4 locuteurs. Nous avons tout d’abord évalué contextuellement le lieu articulatoire de la consonne au moyen de la seule position de la zone de bruit sur l’explosion. La prise en compte des transitions significatives - en association avec le burst - a constitué l’objet d’un second test visant à mesurer si ces deux types de connaissances étaient complémentaires.

3.1. Résultats sur l’explosion
Les résultats obtenus avec les règles caractérisant le lieu d’articulation sur l’explosion de la consonne sont donnés par la matrice de confusion de la table 3. Les performances sont intéressantes pour /i/ et /j/ mais demeurent insuffisantes pour /p/. Les confusions pour la consonne bilabiale résultent d’une absence fréquente du burst et de la diffusion de l’énergie dans le spectre.

Table 3 - Matrice de confusion pour l’identification du lieu articulatoire des oscillations soudées à partir de l’explosion.

4. CONCLUSION
L’identification descendante (contexte phonétique connu) des consonnes occlusives soudées en reconnaissance de la parole est une opération qui peut être effectuée avec de bonnes performances en utilisant des systèmes de représentation des connaissances. Ces techniques ont produit des résultats intéressants dans d’autres circonstances [6] et sont opérationnelles pour la caractérisation multimédia et la discrimination d’autres phénomènes dans des contextes connus ou hypothétiques. La modélisation par auto-organisation des informations de ce type avec un processus d’apprentissage implique la prise en compte d’une grande quantité d’exemples de nombreux locuteurs. Nous envisageons, pour comparer les performances de notre méthode, de réaliser un système utilisant des techniques connexionnistes qui serait supervisé par des règles de manière à fournir des entrées prétraitées aux organes effectuant l’apprentissage et la reconnaissance et limiter ainsi le nombre des exemples nécessaires.

5. REFERENCES
dent recognition of small vocabularies", Speech Communication, volume 10 n°2
Automatic Formant Estimation in a Speech Recognition System

O. Schmidbauer
Siemens AG, ZFE IS KOM31
Otto-Hahn-Ring 6, D-8000 München 83

Abstract
We present an algorithm for formant estimation in continuous speech which is designed to work under "online" conditions in a speech recognition system. The algorithm combines heuristic knowledge about the spectral and temporal behaviour of formants in speech. Preclassification into broad phonetic categories allows to use different algorithms for formant estimation in vowel- and consonant-like regions of speech. Recognition experiments show that formant parameters are a powerful feature set for speech recognition and can compete with other standard feature vectors.

1 Introduction
Formants appear as prominent peaks in the short-time spectra of speech and are defined as the characteristic resonance frequencies of the vocal tract ordered by frequency. Formants carry important information about acoustic-articulatory relations, because they change their frequency and amplitude values according to different vocal tract shapes. They can be viewed as an important source of information in acoustic-phonetic decoding. Thus formants have become a standard in phonetics for describing complex acoustic-phonetic relations.

Formants also seem to be an ideal parameter set for speech recognition, but so far they have not become a standard in this area. The reason is, that automatic formant extraction is not a trivial problem. Already existing algorithms for automatic formant extraction, e.g. [1], [3] show the evidence that formant extraction without any errors is impossible. The significance of information carried by formants is revealed by severe recognition errors in the case of incorrect formant estimation.

The next chapter briefly introduces the problem of automatic formant extraction. Then the different parts of the algorithm are presented. Finally some speech recognition experiments with formant parameters are described.

2 The Problem
Contrary to commonly used feature sets in speech recognition, formants are not defined by a mathematical model, which allows to calculate them directly from the speech wave. They are defined by articulatory phonetics as vocal tract resonances. Formants only can be calculated indirectly via peaks or roots of the power spectrum.

![Figure 1: Formant estimation problem.](image)

In terms of estimation theory, we can therefore formulate the following problem (see also Figure 1): Suppose the peaks and roots $f_1$ (also called "formant candidates") of the power spectrum are the only data which can be measured and which give us some information about the unknown quantity "formants" $f_0$ inside the system. So, depending on $f_1$ we have to make an estimate for the formants $f_0(f_1)$ that the estimation error $E = f_0(f_1) \setminus f_0$ is "small".

However, this estimation process is heavily influenced by two different noise sources $e_0$ and $e_1$. The errors caused by $e_1$ have their origin in the articulatory system. The formant order may be confused by zeros in the vocal tract transfer function. Thus some formants are highly damped and are not detectable. Noise source $e_0$ causes measurement errors, e.g., as the fundamental frequency is superseded to the short-time spectra, prominent pitch peaks may be confused with formant candidates.

Existing methods for automatic formant estimation simply try to map these measured and noisy peaks or roots to formants by temporal smoothness criteria, e.g., [1],[3]. The background for these procedures is the assumption that, due to the inertia of the articulators, the temporal behaviour of real vocal tract resonances (=formants) is indicated by continuity.

The algorithm we present in this paper does not exclusively use smoothness criteria, because this is an over-simplification; we will illustrate this point by two examples; firstly, imagine a vowel-segment where a highly damped formant is missing, smoothness criteria do not help at all to classify the measured peaks into formants; secondly, smoothness criteria may lead to crucial errors at places where formants jump significantly in frequency, tracks of different formants may be connected with each other.

3 The Algorithm
Analyzing carefully the temporal and spectral behaviour of formants in speech and also the nature of possible errors we designed an algorithm which can be divided into four steps (see also figure 2): (1) spectral analysis and preclassification into broad categories of manner of articulation, (2) formant identification (FID) in vowel-like segments without smoothness criteria, (3) formant tracking (FTR) in vowel-consonant (VC) and consonant-vowel (CV) segments with smoothness criteria and (4) preparation and normalization of formant parameters for speech recognition.

![Figure 2: Schematic flow graph for formant extraction.](image)

The algorithm uses 128-point FFT-spectra with a bandwidth of 8kHz. The spectra are calculated via a 16-th order LPC-analysis with a 20ms Hamming window, which is shifted in 10ms steps. The formant candidates are determined both by peak-picking and root solving.

3.3 Preclassification
Initially, the speech signal is preclassified into 7 broad phonetic categories (silence, weak fricative, strong fricative, voiced plosive, nasal, sonorant, vowel) which correspond to manner of articulation. This is due to the assumption, that there is no overlap of formant frequencies in segments with constant manner of articulation. This makes the following steps of the presented procedure, especially step 2 formant identification, more easily. Classification into categories of manner of articulation is performed by mixture density Hidden Markov Models (CDHMM) similar to [4], using very simple acoustic features like energy contour, zero-crossings rate, low frequency energy (up to 1000 Hz) and the ratio of high to low frequency energy.

3.2 Formant Identification
Formants are extracted in vowel-like (V) segments first, because they usually are more prominent in vowels than in consonants and therefore may be detected more easily. The main task of this step is to allocate formant candidates to formants, taking into account that formants may be missing over the whole duration of a V-segment (see also the example in figure 2). $M_F$ formant candidates are calculated every $10\text{ms}$; $M_F$ is set to the number of LPC-roots minus one. Formant identification first tries to find the dominant formant region within a segment. This is accomplished by approximating the distribution of formant candidates in V-segments by $M_F$ cluster centers with gaussian distributions. The procedure itself consists of three steps: (1) initialization of the cluster procedure, (2) classification of cluster centers by k-mean clustering and (3) classification of the formant candidates into formants by a mean square estimator.

(1) Initialization: To initialize the segment specific formant clusters, we first calculate the mean $m_N$ and variance $q_N$ of the formant candidate frequencies $s_N$ over all $N$ frames $l$ of a V-segment:
3.3 Formant Tracking

This part of the algorithm continues the formants of the vowel-like (V) segments into neighboring consonant-like (C) speech segments, i.e., formant tracking works on CV- and VC-segments. The CV- and VC-segments are well defined by preclassification. As the formants of the V-region are already known, this part of the algorithm has the task to correct and complete computed formant tracks by smoothness criteria (see example in figure 2). A non-linear smoothing algorithm based on dynamic programming was chosen for this task. This method is able to keep frequency jumps in some formants by optimizing the overall smoothness of the formant tracks.

The smoothness of the trajectory of formant $f_1$ is measured by a cost function $c_1(l, i, h, f)$. It measures the deviation of formant candidates to the trajectory of formant $f_1$. Assuming that the formant candidates $f_0$ in frame $i$ and $f_0$ belong to the trajectory of formant $f_1$, the costs are given by:

$$c_1(l, i, h, f) = \left| \left( x_{f_0} - x_{f_0} \right) \cdot C_i + (i - h) \cdot C_f \right|$$

with $h = 1, \ldots, 3$, $l = 1, \ldots, M_{f_1}$, $C_i$ and $C_f$ being constants.

The cost function consists of three main terms: The first term corresponds to the frequency distance in Hz, the second term measures the temporal distance between the formant candidates and the third one is a weighting term which corresponds to the reverse probability of the formant candidates belonging to formant $f_1$. The function accepts small values for smooth and large values for corrupted trajectories.

The optimization criterion for the allocation of formant candidates to formants is given by the next formula. The criterion states that the total error $E$ given by the sum of the costs over all frames $N_{VC}$ for a VC- and $N_{CV}$ for a CV-segment respectively, has to be a minimum:

$$E = \sum_{i} \sum_{k} c(l, i, h, f)$$

for all $l, k$ and $f$.

This equation can be elegantly solved by dynamic programming. A solution for this problem is presented in [6].

3.4 Formant Parameters

The formant parameter set which is used for speech recognition consists of 7 formant frequencies and of two energy terms for each formant (a total of 21 parameters). The energy terms correspond to the logarithmic power which is contained in the frequency region expanding from a formant center to the left $m_l$ or the right minimum $m_r$ in the spectrum. With $e(x)$ being the log. power at frequency $x$, the energy to the left and right side $f_{q}$ of a formant center is calculated by:

$$f_{q} = \int_{m_l}^{m_r} e(x) dx$$

All formant parameters are finally normalized to the speakers mean values and variances. With $f_{p}(l)$ now being one of the 21 formant parameters at time $i$ and $m_{p}$ and $\sigma_{p}$ being the speaker specific means and variances of these formant parameters, the normalized formant parameters $f_{n}(l)$ are calculated by:

$$f_{n}(l) = \frac{f_{p}(l) - m_{p}}{\sigma_{p}}$$

Expressed in filter bank terminology: The resulting parameters which are used for speech recognition are filter bank coefficients, where the filter channels have variable center frequencies and bandwidths.

4 Experimental Results

The presented algorithm for automatic formant extraction was tested with speech material of 3 speakers (each with 2 versions of 100 phonetically balanced sentences, i.e. about 10 minutes of continuously spoken speech per speaker). The extracted formant parameters were used for classifying the speech signal into 14 categories of place of articulation (silence, glottal, velar, palatal, alveolar, dental-alveolar, labio-dental, bilabial, labial, labio-velar, labio-alveolar, velar, alveolar, dental). For each of these categories a 16-component cepstral feature vector was used instead of a 64-component feature vector as it is used in [5].

The overall mean recognition rate over three speakers (two male, one female) for 21 formant parameters is 74.9 % for the cepstrum 67.4 % and for the mel-spectrum difference vector 78.5 %.

The results show that the formant vector outperforms the cepstral vector (about 7 % better). The recognition performance compared to the 64-component vector is about 4 percent lower, but it has to be noted that the dimensionality of the formant vector is three times lower than for the 64-component vector and that no temporal context was considered for classification.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>21 formant parameters</th>
<th>16 cepstral coefficients</th>
<th>64 mel differential coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>male1</td>
<td>74.7 / 84.9</td>
<td>66.8 / 80.3</td>
<td>78.4 / 86.7</td>
</tr>
<tr>
<td>male2</td>
<td>74.7 / 84.1</td>
<td>67.3 / 79.5</td>
<td>78.0 / 86.7</td>
</tr>
<tr>
<td>female</td>
<td>75.9 / 86.0</td>
<td>68.1 / 81.1</td>
<td>79.2 / 87.9</td>
</tr>
</tbody>
</table>

Table 2: Form recognition rates [%] for different speakers and different feature sets.

References


REMERCIEMENTS / ACKNOWLEDGEMENTS

Le XIIème Congrès International des Sciences Phonétiques a été organisé avec l'aide de /

The Organisation of the XIIth International Congress of Phonetic Sciences has been supported by

- Centre National de la Recherche Scientifique :
  - Département des Sciences de l'Homme et de la Société
  - Département des Sciences de l'Ingénieur
  - GRECO-PRC Communication Homme-Machine, pôle parole
- Conseil Général des Bouches-du-Rhône
- Conseil Régional, Provence-Alpes-Côte d'Azur
- Délégation Générale à la Langue Française
- Ministère des Affaires Etrangères, Secrétariat d'État aux Relations Culturelles Internationales
- Ministère de l'Education Nationale, de la Jeunesse et des Sports :
  - Direction de la Recherche et des Études Doctorales
  - Direction des Affaires Générales, Internationales et de la Coopération
- Ministère de la Francophonie, Agence de coopération culturelle et technique
- Ministère de la Recherche et de la Technologie
  - Délégation à l'Information Scientifique et Technique
  - Délégation aux Affaires Internationales (Programme ACCES)
- Municipalité d'Aix-en-Provence
- Université de Provence
INDEX DES AUTEURS

INDEX OF AUTHORS

(Volume 4)
TABLE DES MATIERES

CONTENTS
SESSIONS ORALES / ORAL SESSIONS

SESSION 1 : Production

1 Palatoglossus activity during VCV utterances containing oral and nasal consonants of Hindi.
   Prakash Dixit 4:2

2 Some acoustic-phonetic parameters of the Lombard effect for the voice trained.
   William Weiss 4:6

3 Dynamic voice quality variations in female speech.
   Inger Karlsson 4:10

4 Temporal modelling of gestures in articulatory assimilation.
   Martin C. Barry 4:14

5 Articulation of prosodic contrasts in French.
   Janet Fletcher, Eric Varikiotis-Bateson 4:18

6 Essai de méthode pour la recherche de l'image centrale : voyelles [i, e, a] du français.
   Annie Pastor 4:22

7 De l'analyse d'une variation de débit dans la chaîne parlée, à la lumière de la cinéradiographie.
   Béatrice Vaxelaire 4:26

8 Phonological organization in bilinguals : evidence from speech error data.
   Cheney Crow 4:30

9 Coordination du geste et de la parole dans la production d'un instrument traditionnel.
   Véronique Berthier, C. Abry, T. Lallouache 4:34

SESSION 2 : Perception

1 Integration of auditory and visual components of articulatory information in the human brain.
   Reijo Aulanko, Mikko Sams 4:38

2 An objective and a subjective approach to speaker recognition.
   Elisabeth Lhote, Laura Abou Haidar 4:42

3 Frequency modulation of formant-like spectral peaks.
   Peter J. Bailey, Kim Bevan 4:46

   Marie-Agnès Caihiard, Guy Tiberghien, A. Cior-Tseva, Tahar Lallouache, Pierre Escudier 4:50

5 Occlusive silence duration of velar stop and voicing perception for normal and hearing-impaired subjects.
   Yves Cazals, Lionel Palis 4:54

6 Perception of syncope in native and non-native American English.
   Joann Fokes, Z.S. Bond 4:58

7 Central mechanisms of vowel. Perception, categorization and imitation.
   Inna Vartanian, Tatiana V. Chernikovskaya 4:62

8 Factors affecting the given-new distinction in speech.
   Sarah Hawkins, Paul Warren 4:66

9 Central mechanisms of intonation processing - comprehension and imitation.
   Tatiana Chernigovskaya, Inna V. Vartanian 4:70

SESSION 3 : Perception

1 L'influence de la durée dans l'identification des liquides : étude comparée en espagnol de Buenos Aires et en français de Montréal.
   Benoît Jacques, Maria Amalia Garcia Jurado, Miguelina Guirao 4:74

2 Perception of anticipatory VCV-coarticulation : effects of vowel context and accent distribution.
   Vincent-Johan Van Heuven, Marc Ch. Dupuis 4:78

3 Effect of vowel quality on pitch perception.
   Ilkka-Mauri Raino, Olli Aaltonen, A. Hellström, E. Vilkman 4:82
<table>
<thead>
<tr>
<th>Session</th>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>The perception of silent-center syllables in noise.</td>
<td>Susan L. Hura</td>
<td>4:86</td>
</tr>
<tr>
<td>5</td>
<td>Minimal duration for perception of full-spectrum vowels.</td>
<td>Rudolf Weiss, W.S. Brown (jr.), Shaw N. Gynan</td>
<td>4:90</td>
</tr>
<tr>
<td>7</td>
<td>Understanding disfluent speech: is there an editing signal?</td>
<td>Robin Lickley, Richard Shillcock, Ellen Garman Bard</td>
<td>4:98</td>
</tr>
<tr>
<td>8</td>
<td>Individual variability in the perception of cues to an initial BA-PA voicing contrast.</td>
<td>Valerie Hazan, Bo Shi</td>
<td>4:102</td>
</tr>
<tr>
<td>9</td>
<td>Perceptual spaces of the Russian vowels.</td>
<td>Galina Lebedeva</td>
<td>4:106</td>
</tr>
<tr>
<td></td>
<td><strong>SESSION 4 : Phonétique descriptive / Descriptive phonetics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>La phonétisation du Castillan.</td>
<td>Claudine Cabrera, M. Contini, L.J. Boë</td>
<td>4:114</td>
</tr>
<tr>
<td>3</td>
<td>Language specific patterns of prosodic and segmental structures in Swedish, French and English.</td>
<td>Gunnar Fant, Anita Kruckenberg, Lennart Nord</td>
<td>4:118</td>
</tr>
<tr>
<td>4</td>
<td>Towards an account of language-specific patterns of the timing of voicing.</td>
<td>Caroline Wills, Gerard Docherty</td>
<td>4:122</td>
</tr>
<tr>
<td>5</td>
<td>Instrumental phonetic fieldwork: techniques and results.</td>
<td>Peter Ladefoged</td>
<td>4:126</td>
</tr>
<tr>
<td>6</td>
<td>An acoustic study of Xhosa Clicks.</td>
<td>Bonny Sands</td>
<td>4:130</td>
</tr>
<tr>
<td>7</td>
<td>The effect of linguistic expectancy on phonetic transcription: developing an adequate alignment algorithm.</td>
<td>Catia Cucchiarini, R. Van Bezooijen</td>
<td>4:134</td>
</tr>
<tr>
<td>8</td>
<td>Phonetic transcription as a means of diagnostically evaluating synthetic speech.</td>
<td>Renée Van Bezooijen, Wilhelm Vieregge</td>
<td>4:138</td>
</tr>
<tr>
<td>9</td>
<td>Consonant clusters: a comparison between word internal and word juncture.</td>
<td>Christine Meunier</td>
<td>4:142</td>
</tr>
</tbody>
</table>

**SESSION 5 : acquisition**

<table>
<thead>
<tr>
<th>Session</th>
<th>Title</th>
<th>Authors</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Speaking while intoxicated: phonetic and forensic aspects</td>
<td>Angelika Braun</td>
<td>4:146</td>
</tr>
<tr>
<td>2</td>
<td>Temporal control in speech of children and adults.</td>
<td>Cécile Kuipers</td>
<td>4:150</td>
</tr>
<tr>
<td>3</td>
<td>Premeaningful vocalizations of hearing-impaired and normally hearing subjects.</td>
<td>Carol StoeI-Gammon</td>
<td>4:154</td>
</tr>
<tr>
<td>4</td>
<td>A longitudinal study of the speech acquisition of three siblings diagnosed as verbally dyspraxic.</td>
<td>Maryanne Tate</td>
<td>4:158</td>
</tr>
<tr>
<td>5</td>
<td>Examination of language-specific influences in infants' discrimination of prosodic categories.</td>
<td>Catherine T. Best, Andrea Leviit, Gerald McRoberts</td>
<td>4:162</td>
</tr>
<tr>
<td>6</td>
<td>Articulatory organization of early words: from syllable to phoneme.</td>
<td>Elisabeth W. Goodell, Michael Studdert-Kennedy</td>
<td>4:166</td>
</tr>
<tr>
<td>7</td>
<td>Rhythmic phenomena in a child's babbling and one-word sentences.</td>
<td>Morio Kohno, T. Tsushima</td>
<td>4:170</td>
</tr>
<tr>
<td>8</td>
<td>L'intonation de question dans le langage émergent.</td>
<td>Gabrielle Konopczynski</td>
<td>4:174</td>
</tr>
</tbody>
</table>
SESSION 6 : prosodie / prosody
Intonation

1 Strategies for prosodic phrasing in Swedish.
Gösta Bruce, Björn Granström, David House 4:182

2 The interaction of fundamental frequency and intensity in the perception of intonation.
Klaus Kohler 4:186

3 Rhythmic patterns of the discourse in Mexican Spanish and Brazilian Portuguese.
Antonia Simões 4:190

4 Syntax and intonation in Italian noun phrases.
Shiro Kori, Hiroko Yasuda 4:194

5 The role of intonation as a marker of semantic associations and enunciative operations in English.
Jennifer Low 4:198

Kikuo Maekawa 4:202

7 Combinations of types of pitch accent in a corpus of Russian speech.
Cecilia Odé 4:206

8 A crosslinguistic description of intonation contours of a multilanguage text-to-speech system.
Gábor Oiaszy 4:210

9 Measuring intonation at low signal-to-noise-ratios.
Vytautas Pikiūna 4:214

SESSION 7 : Prosodie / Prosody

1 Speech F0 extraction based on Licklider's pitch perception model.
Alain De Cheveigne 4:218

2 A computer assisted method of investigating intonational correlations in adjacent utterances.
Christian Sappok, Rozalija Kasatkina, Sandro Kodzasov 4:222

3 The beginning of Germanic prosody.
Anatoly Liberman 4:226

4 Timing in Catalan.
Daniel Recasens 4:230

5 The timing of vowel and consonant gestures in Italian and Japanese.
Caroline Smith 4:234

6 Pausing in texts read aloud.
Eva Strangert 4:238

7 Rhythmic structures in poetry reading.
Anita Kruckenberg, Gunnar F. Fant, Lennart Nord 4:242

8 A contrastive analysis of Spanish and Catalan Rhythm.
Mar Carrio i Font, Antonio Rios Mestre 4:246

9 Rhythmic model of a phonetical word of present-day Lithuanian utterances.
Lilia Anusiene 4:250

SESSION 8 :

1 Incidences du trait phonologique de durée vocalique sur la prosodie du français québécois.
Laurent Santerre 4:254

2 Perceiving rhythm in French ?
Jacqueline Vaissière 4:258

3 Tone production in standard Chinese : EMG data and command-response modelling.
Pierre Hallé 4:262

4 Acoustic correlates of stress shift.
Stefanie Shattuck-Hufnagel 4:266

5 Terminality and completion in Danish, Swedish and German.
Nina Gronnum 4:270

6 Intonation modelling in a text generation program.
Carlos Gussenhoven, Toni Rietveld, Lex Elich 4:274
Ways of exploring speaker characteristics and speaking styles.  
Björn Granström, Lennart Nord  
4:278

Analyse de la prosodie de la parole spontanée en suédois et en français.  
Paul Touati  
4:282

Intonation patterns in Greek discourse.  
Antonis Botinis  
4:286

SESSION 9 : Phonétique appliquée : Applied phonetics

1 The most important difficulties when teaching Spanish phonetics to Czechs.  
Jana Kullova  
4:290

2 Aspects of the relation between intonation and the interpretation of poems.  
Johann Uys  
4:294

3 Effects of voice characteristics on attitude change.  
Claire Gélinas-Chebat, Jean-Charles Chebat  
4:298

4 Parole chantée et parole déclamée : autour de Salomé.  
Aspects articulatoires, rythmiques etintonatifs.  
Sibylle Vater  
4:302

5 Phonostylistics in foreign language learning.  
Kathryn Keller  
4:306

6 Riez-vous en hi ! hi ! hi ! ou en ah ! ah ! ah ! oh ! oh !  
Pierre Léon  
4:310

7 Variables intonatives chez la femme vénézuélienne.  
Elsa Mora  
4:314

8 Étude des paramètres temporels des voix sans larynx.  
Lise Crevier-Buchman  
4:318

9 Phonetic aspects of speech produced without a larynx.  
Lennart Nord, Britta Hammarberg, Elisabet Lundström  
4:322

SESSION 10 : Pathologie / Pathology

1 On using intensity as a coding parameter in tactile speech stimuli : psychophysiological discriminability effects.  
Hans Georg Piroth, Stephan Gfroerer  
4:326

2 Speech perception abilities of patients using cochlear implants, vibrotactile aids and hearing aids.  
Eva Agelfors, Arne Risberg  
4:330

3 Changes in speech breathing following cochlear implant in postlingually deafened adults.  
Harlan Lane, Joseph Perkell, Mario Svirsky, Jane Webster  
4:334

4 Compensatory articulation and nasal emission of air in cleft palate speech with special reference to the reinforcement theory.  
Birgit Elisabeth Hutters, Kirsten Brondsted  
4:338

5 Perceptual and acoustic analysis of the voice in acute laryngitis.  
Anders Lôfqvist, Lucyna Schalén  
4:342

6 The development of articulatory skills in cleft palate babies.  
Kino Jansonius-Schultheiss  
4:346

7 Acoustic evidence that postlingually acquired deafness affects speech production.  
Roddy Cowie, Ellen Douglas-Cowie, Joan Rahilly  
4:350

8 Mean-term perturbations of the pseudo-period of the glottal waveform.  
Raoul De Guchteneere, Jean Schoentgen  
4:354

9 The interaction of speech perception and reading ability.  
Mária Gósy  
4:358
SESSIONS AFFICHEES / POSTER SESSIONS

SESSION 11 : Acoustique / Acoustics

1 Analysis of glottal waveform in different phonation types using the new IAIF-method.
   Paavo Alku, Erkki Vilkmn, Unto K. Laine.............. 4:362

2 A voice conversion method and its application to pathological voices.
   Hisao Kuwabara, Tokru Takagi.......................... 4:366

3 Consistency in /r/ trajectories in American English.
   Carol Espy-Wilson...................................... 4:370

4 La variabilité inter-locuteur, étude sur les réalisations acoustiques de /e,ɛ/.
   Anne Bonneau............................................. 4:374

5 Some SARA vowel inventories and vowel system predictions.
   Djita Issa Djarangar..................................... 4:378

6 Locus-nucleus relation and vot in spontaneous and elicited speech.
   Diana Krull................................................. 4:382

7 A study of [r] and [ɾ] in spontaneous speech.
   Carme de la Mota Gorriz.................................. 4:386

   Toomas Alliosaar, Matti Karjalainen.................... 4:390

9 Bulgarian vowel clusters and statistics by 30 male and 30 female speakers.
   Philip Christov.............................................. 4:394

SESSION 12 : Aspects linguistiques / Linguistic aspects

1 Phonology of synharmonism and a new synharmonic script.
   Alimchan Dzhunisbekov.................................. 4:398

2 De l'indépendance du phonème faible au système phonologique de la langue russe.
   Svetlana Nikolaienva Dmitrenko........................ 4:402

3 Modelling vowel systems by effort and contrast.
   Louis Ten Bosch............................................ 4:406

4 Articulatory and perceptive aspects of typology of sound systems in conditions of multilingualism.
   Tatiana Verbitskaya, T. Korolyova...................... 4:410

5 The influence of social factors on urban speech.
   Lidia Ignatkina............................................. 4:414

6 Regional voice quality variation in Sweden.
   Claes-Christian Elert, Britta Hammarberg............ 4:418

7 Etude sur la perception de l"accent" régional du nord et de l'est de la France.
   Fernand Carton, Robert Espesser, Jacqueline Vaissière 4:422 X

8 The effect of addressee familiarity on word duration.
   Janice McAllister, Cathy Sotillo, Ellen Gurman Bard 4:426

9 Computer data base and orthoepic studies.
   Natalia Bogdanova.......................................... 4:430

10 Sur la classification universelle des sons du langage et l'Aphi.
    Myrrha Gordina............................................ 4:434

11 Gémination phonétique en frontière de mots.
    Alain Marchal, A.S. Del Negro.......................... 4:438

12 Consonant clusters and their connection with the morphological structure of the Kazakh word.
    Svetlana Meirbekova...................................... 4:442

13 Understanding "hm", "mhm", "mmh".
    Stefan Werner............................................ 4:446

14 Age effects in acquisition of non-native phonemes : perception of English /r/ and /l/ for native speakers of Japanese.
    Yamada Reiko, Yoh'ichi Tohkura......................... 4:450
Andrea Levitt, Qi Wang 4:454

16 The unintelligibility of speech to children : effects of referent availability.
Ellen Gurman Bard, Anne H. Anderson 4:458

17 On the phonetic system evolution in some archaic Russian Dialects.
Sergey Knyazev, Sofia Pojaritskaya 4:462

SESSION 13 : Technologie / Technology
Reconnaissance automatique de la parole / Automatic speech recognition

1 Modelling articulatory inter-timing variation in a speech recognition system.
Mats Blomberg 4:466

2 Including duration information in a threshold-based rejector for hmm speech recognition.
Antonio M. Peinado, Antonio J. Rubio, Juan M. Lopez, José C. Segura, Victoria E. Sanchez 4:470

3 Modelization of allophones in a speech recognition system.
Katarina Bartkova, Denis Jouvet 4:474

4 Towards more reliable automatic recognition of the phonetic units.
Arvydas Domatas, Algimantas Rudzionis 4:478

5 The SYLK project : syllable structures as a basis for evidential reasoning with phonetic knowledge.
Peter-John Roach, D. Miller, P.D. Green, Antony J. Simons 4:482

6 Recovering tube kinematics using time-varying acoustic information.
Richard Mc Gowan 4:486

7 Phoneme-like model of speech signal.
Stasys Noreika, Algimantas Rudzionis 4:490

8 Représentation de connaissances indépendantes du locuteur pour la reconnaissance de mots acoustiquement proches.
Henri Meloni, Philippe Gilles 4:494

9 Automatic formant estimation in a speech recognition system.
Otto Schmidbauer 4:498