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 [8, 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 vowels, and by Dixit et al [4] on Hindi front nasal vowels provided unequivocal support for the "gate-pull" model of nasal sound production. The PG data reported by Dixit et 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 perorally 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 /t d n/, and V represented /i 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/, /utu/, /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 vowels 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 even 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 also. 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 Figure 1.

In this figure, 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 vis-a-vis 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/, /utu/, /ada/ and /udu/ containing an oral consonant surrounded by the back oral vowels. Both these peaks appear to be

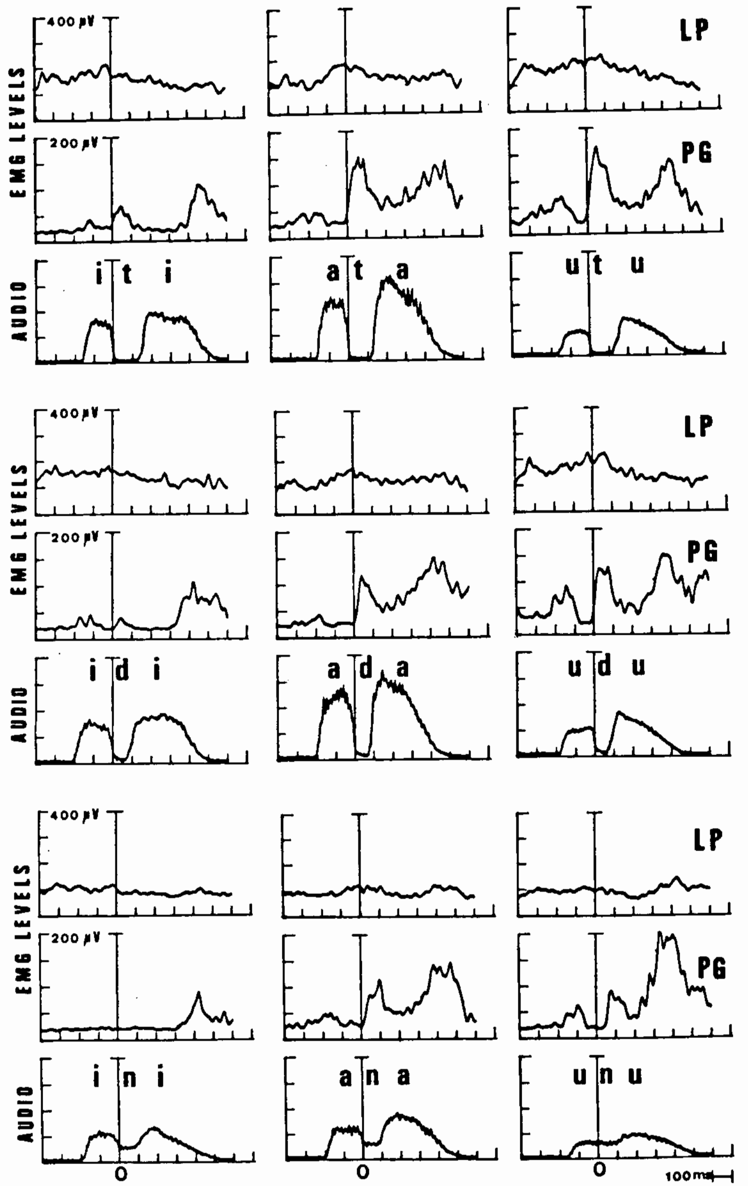


Fig. 1 Superimposed curves of ensemble averages of LP and PG EMG signals and audio signals for the experimental utterances. Audio and EMG signal amplitudes in arbitrary units and microvolts, respectively, are represented along the ordinate. The units along the absicssa represent 100 ms intervals. Zero (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 [9]). 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 uV) 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 relat
ed 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

[1] BELL-BERTI, F. (1976), "An electromyographic study of velopharyngeal function in speech," J. SpeechHear. Res., 19, 225-240.

[2] BELL-BERTI, F. & HIROSE, H. (1973), "Patterns of palatoglossus activity and their implications for speech organization," Haskins Lab. Status Rep. Speech Res., SR 34, 203-209.

[3] BENGUEREL, A.P.; HIROSE, H.; SAWASHIMA, M. & USHIJIMA, T. (1977), "Velar coarticulation in French: an electromyographic study", J. Phonet., 5, 159-167.

[4] DIXIT, R.P.; BELL-BERTI, F. & HARRIS, K.S. (1987), "Palatoglossus activity during nasal/nonnasal vowels of Hindi," Phonetica, 44, 210-226.

[5] FRITZELL, B. (1969) "The velopharyngeal muscles in speech: an electromyographic and cineradiographic study", Acta Oto-Lar., Suppl. 250.

[6] KUEHN, D.P.; FOLKINS, J.W. & CUTTING, C.B. (1982), "Relation- ship between muscle activity and velar position", Cleft Palate J., 7, 25-35.

[7] LUBKER, J.; FRITZELL, B. & LINDQVIST, J. (1970), "Velopharyngeal function in speech: an electromyographic study", Q. Prog. Status Rep., Speech Transm. Lab., R. Inst. Technol., Stockholm, No. 4, 9-20.

[8] LUBKER, J.; LINDQVIST, J. & FRITZELL, B. (1972), "Some temporal characteristics of velopharyngeal muscle function," in Phonetic Symposium (University of Essex Language Center, Essex), 226-268.

[9] LUBKER, J & MAY, K (1973), "Palatoglossus function in normal speech production," Papers from the Inst. Ling., Univ. Stokh. (PILUS), 17, 17-26.

SOME ACOUSTIC-PHONETIC PARAMETERS OF THE LOMBARD EFFECT FOR THE VOICE TRAINED

William Weiss

University of Ottawa, Ottawa, Ontario K1N 6N5, Canada.

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:1.Voice Trained subjects increase vocal levels less than Untrained subjects under masked conditions, therefore showing an attenuated Lombard effect. 2. Some reported voice quality measurements (1.αAB= >1000Hz / <1000Hz, 2.θ=F1 / F0) do not seem to apply to speech of Francophones.

1. INTRODUCTION

It is well known that the presence of noise produces an increase in vocal levels ([3]Lombard, 1911; [2]Lane and Tranel, 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 singers' performance deteriorates less than that of amateur musicians ([6]Ward & Burns, 1978). That is attributed to a process of kinesthetization, 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 quastions 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?

2. METHOD

2.1. Voice quality measurements

An acoustic measure of voice quality was proposed by [1]Frokjaer-Jensen and Prytz (1976) as α=intensity above 1kHz / intensity below 1kHz. [7]Wedin et al. (1978) seemed to confirm the utility of this measure in speech with a group that had undergone voice training. [5]Sundberg and Gauffin (1978), seemed to suggest that in singing, judging the higher spectra as a measure of good quality is misleading because it could be obtained with an increased vocal effort ("pressed" phonation) which is not characteristic of trained male singers. They proposed that a measure of good quality is a higher increase of energy in the F0 area relative to the F1 area of trained subjects ("flow" phonation). In order to utilize these voice quality acoustic measurements, this experiment extracted Long Term Average Spectra for the following intervals:

F0e:Log energy at interval 80-160Hz for men, 160-250Hz for women

F1e: Log energy at interval 315-600Hz

B1K: Log energy below 1kHz (80-800Hz)

A1K: Log energy above 1kHz (1000-5000Hz)

θF1F0: F1e minus F0e

αAB= A1K minus B1K

These intervals also served to compare spectral levels.

2.2. Subjects

The group of 23 subjects was composed of 1. Voice Trained (n=12) and Untrained subjects (n=11). Subjects with abnormal hearing or with mother tongues other than Canadian French (Francophones) were excluded. The trained subjects were either members of a well known choir or professional actors and radio announcers. The subjects donated their time without pay.

2.3. Materials

The French text, of phonetically balanced contents lasting approximately one minute of reading time, was edited from existing literary materials.

2.4. Procedure

The subjects were recorded while reading the same one minute text under three conditions: 1.Normal reading (S); 2.With right ear masked with a 75dB white noise(SRM); 3.With left ear masked with a 75dB white noise(SLM).

All the recordings, and the audiometric screening, were conducted in a soundproof cabin (I.A.C.). The microphone was a Sennheiser MD441-U (filtration switch on 'M'), the tape recorder a full track Revox 77A (tape speed 15 ips), and the tapes Ampex 406.

The masking noise was produced with the Maico Precision Hearing Test Instrument MA-24, through Maico headphones with one earphone removed. In the conditions of masking, the subjects had one ear masked with noise whereas the other remained free. This procedure was adopted for future analysis of laterality effects. The recordings were performed at one foot distance from the microphone and the order of the three conditions was systematicaly varied for succeeding subjects.

2.5. Analyses

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 transfered and digitized in an IBM microcomputer through a software package designed for the project and then transfered to the mainframe computer where Spectral levels were determined for each of the three recording conditions.

3. RESULTS

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

Speech condition		F0e	F1e	θF1F0	B1K	A1K	αAB
		Interval dB					
Normal Speech(S)	Tr.Fra.:	-21.73	-22.14	-0.41	-17.93	-29.61	-11.67
	Untr.Fra.:	-20.96	-20.48	0.47	-16.53	-28.04	-11.51
		**	***		***	**	
Speech with right ear masked (SRM)	T:	-21.06	-19.86	1.20	-16.60	-26.64	-10.03
	U:	-18.28	-15.60	2.67	-12.66	-21.90	-9.24
		*	**		**	*	
Speech with left ear masked (SLM)	T:	-21.81	-20.84	0.97	-17.45	-27.29	-9.83
	U:	-18.54	-16.71	2.14	-13.62	-23.60	-9.98
		*	*		**		
S-SRM (R)	T:	-0.66	-2.28	-1.62	-1.32	-2.97	-1.64
	U:	-2.68	-4.88	-2.19	-3.87	-6.14	-2.26
		**	*		**		
S-SLM (L)	T:	0.08	-1.30	-1.38	-0.47	-2.32	-1.84
	U:	-2.11	-3.77	-1.66	-2.91	-4.43	-1.52

* significant at the 0.05 level
** significant at the 0.01 level
*** significant at the 0.001 level
F0e:Energy at interval 80-160Hz for men, 160-250Hz for women
F1e:Energy at interval 315-600Hz
B1K:Energy below 800Hz (80-800Hz in 1/3 octaves)
A1K:Energy above 1000Hz (1000-5000Hz in 1/3 octaves)

The table above shows the following results:

1. There are no significant differences in the Normal Speech condition for spectral levels (F0e,F1e, B1K,A1K) and voice quality (θF1F0, αAB) between trained and untrained subjects.
2. Spectral levels of voice trained subjects are significantly lower in both masked conditions (For SRM: F0e,$p<.01$; F1,$p<.0006$; B1K, $p<.0005$; A1K,$p<.002$; for SLM: F0e,$p<.02$; F1,$p<.002$; B1K,$p<.002$; A1K,$p<.04$).
3. There are no significant voice quality differences (θF1F0, αAB) 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 quality measurement αAB 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 measurement, θF1F0, 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

[1] Frokjaer-Jensen,B.,and Prytz, S. (1976). "Registration of voice quality," Bruel & Kjaer, Technical Revue, 3, 3-17.
[2] Lane, H.and Tranel, B. (1971). "The Lombard Sign and the Role of Hearing in Speech," J.Speech Hear.Res.14,677-709.
[3] Lombard, E. (1911). "Le signe de l'elevation de la voix," Annales des Maladies de l'Oreille et du Larynx, 37,101-119.
[4] Pick Jr.,H.L., Siegel, G. M.,Fox, P.W., Garber, S.R., Kearney, J.K.(1989). "Inhibiting the Lombard Effect," J. Acoust. Soc. Am. 85,2, 894-900.
[5] Sundberg,J.,and Gauffin,J.(1978)."Waveforms and spectrum of the glottal voice source," Speech Transmis. Lab. Q. Prog. Stat. Rep. STL-QPSR 2-3, Royal Institute of Technology, Stockholm, Sweden, 35-50.
[6] Ward,W.D., and Burns, E.M.(1978)."Singing without auditory feedback,"J. of Res. in Singing,1,2,24-44.
[7] Wedin ,S., Leanderson,R.,and Wedin, L. (1978). "Evaluation of Voice Training.Spectral Analysis Compared with Listeners' Judgements," Folia Phoniatrica,30,103-112.

DYNAMIC VOICE QUALITY VARIATIONS IN FEMALE SPEECH

Inger Karlsson

Dept of Speech Communication and Music Acoustics, KTH, Box 70014, S 100 44 Stockholm, Sweden

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 have 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 large 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 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 concentrated on the female 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 fitting 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 F0. RK corresponds to 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, F0, 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. In a transition between a vowel and an [l] 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 F0 in sentences uttered by different speakers. The correlation coefficient was found to be in the order of 0.75. Deviations occurred for F0 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-

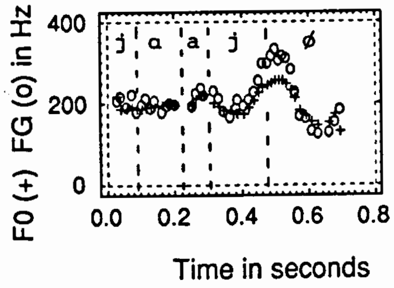


Figure 1. Typical covariation of F0 and FG=F0·RG/100 for a short utterance.

pulse variation. This is due to the uncertainty in defining the exact time for opening. Another source of error is formant frequency differences in the open and the closed interval of the voice pulse. In inverse filtering only one formant value was used in each period. This will result in incomplete cancelling of formant ringings in the open phase and some lack of precision in determining the point of maximum flow. All RK values discussed in this paper are average values, which should minimize these errors.

3. VOICE SOURCE 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 [l, r] 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 unstressed vowel. 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 pole/zero pair, especially in [n], [l] and [v].

Voice source parameters for consonants and for some unstressed vowels in the same sentences are given in Table 1. The values are averaged over at least 4 period. 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, j, n, h] 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], [h] and for one speaker [j]. A possible reason for the high FA is discussed below under "Noise excitation". FA was only slightly higher than F0 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 unstressed vowels for two female speakers. The last column gives the number of occurrences of the phoneme in the investigated sentences. F0 and FA are given in Hz and RK and RG in percent. EE is given in uncalibrated dB so only comparisons within a speaker is possible.

speaker W1			consonants			
	F0	EE	RK	RG	FA	n
[v]	185	50	48	109	215	1
[l]	191	55	48	116	356	8
[n]	270	53	51	141	297	1
[r]	253	57	42	120	657	1
[j]	161	56	43	101	849	1
[h]	216	57	51	107	492	1
stop	232	49	58	101	354	3
unstressed vowels						
[a]	212	61	45	124	867	4
[I]	218	55	38	104	500	2

speaker W2			consonants			
	F0	EE	RK	RG	FA	no
[v]	237	48	59	119	362	2
[l]	209	57	48	111	322	8
[n]	186	56	37	104	208	1
[r]	222	57	43	115	678	3
[j]	196	57	43	100	255	1
[h]	250	57	72	123	600	1
stop	220	51	66	112	343	2
unstressed vowels						
[a]	213	62	35	96	763	4
[I]	277	60	28	90	480	1

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: [l] and [n]. For [l] and [n] the zero and the connected pole are normally due to the geometry of the vocal tract. Zero/pole pairs found in [l] and [n] for two female speakers are given in Table 2.

The zero sometimes detected in [v] as well as a low zero, about 900 Hz, sometimes found in [l], 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. These 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 cancelling 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 a Rothenberg 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 [h] 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

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 BPn its bandwidth. All are given in Hz. * denotes presumed voice source zeros and poles.

		Z1	BZ1	Z2	BZ2	P1	BP1	P2	BP2
W1	[l]	*940	300	2050	300	*1450	150	2500	200
W2	[l]	*990	450	2200	300	*1600	200	2550	200
W1	[n]	750	450			1900	100		
W2	[n]	860	150			1700	100		
W1	[v]	*700	250			*1650	150		
W2	[v]	*840	350			*1600	200		
W1	[h]	1850	600			3000	500		
W2	[h]	2300	600			2600	450		

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 Telecom

7. REFERENCES

[1] FANT, G., LILJENCRANTS, J & LIN, Q. (1985): "A four-parameter model of glottal flow", *STL/QPSR 4/85*, 1-13

[2] GOBL, C. & KARLSSON, I. (1989): "Male and female voice source dynamics." *Proc. of Vocal Fold Physiology Conference, Stockholm.*

[3] KARLSSON I. (1990): "Voice source dynamics for female speakers" *Proc. of the 1990 Int. Conf. on Spoken Language Processing, Kobe.* 69-72

[4] KLATT, D. & KLATT, L. (1990): "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *JASA 87*, 820-857.

[5] ROTHENBERG M (1973): " A new inverse filtering technique for deriving the glottal air flow waveform during voicing", *JASA 53*, 1632-1645

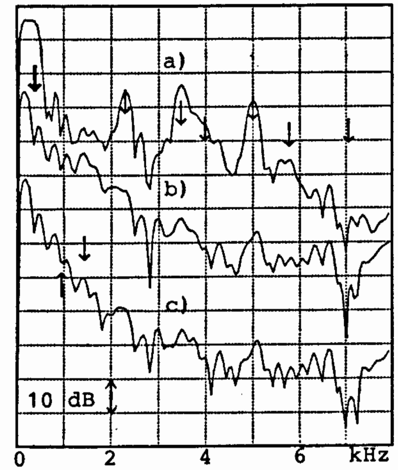


Figure 2. Spectra for one fundamental period of a vowel before and after inverse filtering. From top to bottom the spectrum of a) an unfiltered voice pulse, b) the same pulse with the formants cancelled and c) with an extra pole/zero pair cancelled as well. The formant anti-filters are marked by down-pointing arrows in the upper part of the figure and the extra pole/zero with arrows in the lower part.

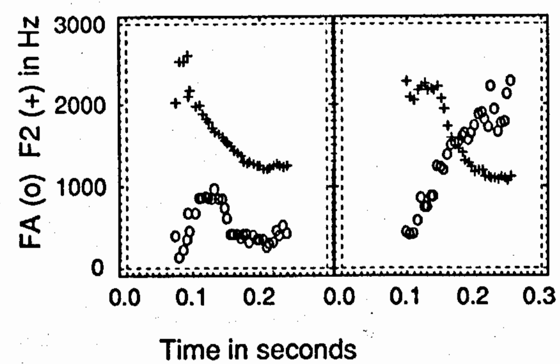


Figure 3. FA variations in a transition from [j] to [ɑ]. F2 is plotted to illustrate the transition. The left half shows a leaky voice, the right part a more sonorant voice.

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 some utterances by the English speakers. The remaining data involve partial loss of the coronal gesture, and are therefore not susceptible to description within conventional phonological formalisms. 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 behaviour 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 and arguments 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 C_1 is an alveolar stop or nasal and C_2 a velar stop, with an intervening mor-

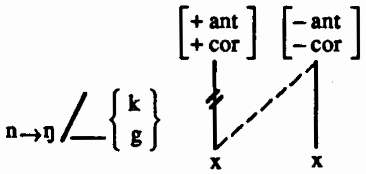


Figure 1: conventional phonological representations for alveolar and dental assimilation

pheme 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 modelled 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.

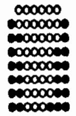


Figure 2: EPG contact pattern for a weakened alveolar gesture

3. QUANTITATIVE INVESTIGATIONS OF ARTICULATORY GESTURES

Further insight into patterns of articulatory activity may be gained by a consideration in terms of the trajectories of individual articulatory subsystems, recently restored to the phonetician's armoury through the development of the concept of the *gesture* in the paradigm of Articulatory Phonology developed by Browman and Goldstein [3]. In the work reported in the present paper gestural trajectories were approximated from time-varying summations of EPG contact

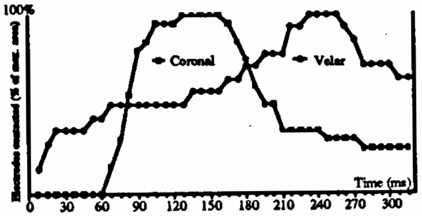


Figure 3: Gestural Trajectories for [n g] in 'hand grenade' (slow utterance)

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 [n g] 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, DDG); (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 onsets of the two closures, or, in the case where no alveolar closure was formed, between the peak in the coronal gesture and the onset of velar closure (INT)[1].

In comparison with the slow utterance, for the fast utterance (fig. 4) CMAX is reduced to 70% of its maximum possible

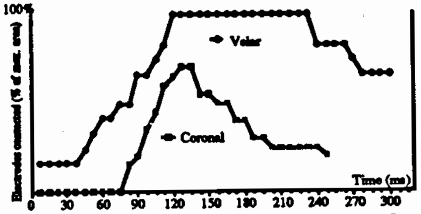


Figure 4: Gestural Trajectories for [n g] in 'hand grenade' (fast utterance)

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 DDG 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 suggest 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 two conso-

[1]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 timing of velic lowering and raising. The nasalisation associated with [n] is retained even when the coronal gesture is lost altogether, giving rise to a velar nasal [ŋ].

nantal timing slots, and the original underlying alveolar place-autosegment is partially delinked.

4. 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 C_1 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 are words such as /sanktsia/ and /funktsia/, in which the nasal and the following stop *must* be syllabified together (since the sequence /kts/ is impermissible as a syllable-onset).[2] These forms showed [ŋ] 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 speech, 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 dorsal gesture (CMAX decreases but DDG remains constant, or even undergoes a slight reduction typical at increased rates of speech), and while INT 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 delinking 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 data. This view accords with Ohala's view [4] that if a 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 incautious, 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 utterances such as these is determined at least in part under the cognitive control of the speaker at least in so far as that the process is seen to apply in some contexts and not others. The fact that the dental 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)
  Expressible in conventional phonological formalisms
  e.g. distribution of Russian [ŋ]; intramorphemic NC clusters in English
- Low-level phonological rules (cf. postlexical rules)
  Partial applications not expressible
  e.g. English alveolar C_1 in CC clusters across morpheme boundaries
- Phonetic effects
  Phonetically motivated articulatory processes; may be phonologically blocked (*e.g. Russian [t,d]*) or may apply freely (*e.g. Russian [n]*)

Two important consequences emerge: that some aspects of the speaker's knowledge of how their language is pronounced involve 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 the variety of low-level phonetic detail which, as I have argued, falls within the subject-matter of a comprehensive theory of phonology. Gestural scores correspond 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 general 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 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 short of 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 'ballistic' 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 existing 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

[1] BARRY, M.C. (1985), A palatographic study of connected speech processes. *Cambridge Papers in Phonetics and Experimental Linguistics*, 4:1-16. (Cambridge University Linguistics Department).

[2] BARRY, M.C. (in prep.), *A cross-linguistic study of connected speech processes.* PhD dissertation, University of Cambridge.

[3] BROWMAN, C.P. & GOLDSTEIN, L. (1990), Tiers in articulatory phonology, with some implications for casual speech. *Papers in Laboratory Phonology I*, J. Kingston & M.E. Beckman, eds. (Cambridge: CUP). pp 341–381.

[4] OHALA, J.J. (forthcoming), Comments on Nolan's 'The descriptive role of segments: evidence from assimilation'. *Papers in Laboratory Phonology II* D.R. Ladd & G.J. Docherty, eds. (Cambridge: CUP).

[2] The principles governing this small class of exceptional forms are discussed further in [2].

ARTICULATION OF PROSODIC CONTRASTS IN FRENCH

J. Fletcher * and E. Vatikiotis-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 non-final), and note that accent-related duration differences are somewhat reduced in the phrase-internal 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 I 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, accented (PAPE), and tonic accented 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 I. Carrier sentences containing the test tokens (in upper case)

Set A	
(i)	Le PAPE a patté Miné.
	Le PAPA pattait Miné.
(ii)	Le PAPE Aballe pattait Miné.
	Le PAPA Bahl pattait Miné.
Set B	
(i)	Miné lechait le PAPE.
	Miné lechait le PAPA.
(ii)	Miné lechait le PAPE?
	Miné lechait le PAPA?

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.

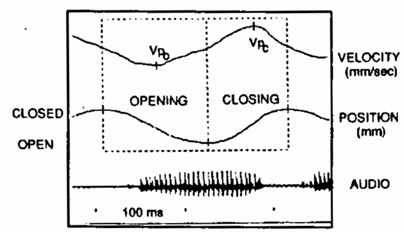


Figure 1
Kinematic Measures

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 /pap/ 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 were 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 61.5, 15.19), with no significant differences in peak velocity. Both acceleration and deceleration duration are 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 tonic/pretonic contrast in /pap/ 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 3.63).

TABLE II - Mean and standard deviation values (in parentheses) of opening gesture duration (ms), lip aperture (mm), peak velocity (mm/s), acceleration and deceleration durations (ms) in /pa/ syllables (token - PAPA).

		Unacc.	Tonic	Final(LOW)	Final(RIS)
D.	AS	101(8)	169(11)	168(6)	182(14)
	BA	73(5)	72(2)	131(11)	135(12)
LA.	AS	7.62(.69)	9.9(1.2)	9.2(1.6)	10.05(.7)
	BA	5.41(.88)	8.31(.83)	9.8(.78)	12.04(1.5)
Vp.	AS	149(13)	152(16)	125(25)	114(20)
	BA	169(30)	134((24)	119(21)	153(35)
Acc.	AS	63(13)	103(8)	84(11)	80(14)
	BA	43(6)	42(4)	75(9)	72(7)
Decel.	AS	39(6)	67(10)	84(15)	103(11)
	BA	30(2)	30(4)	55(6)	64(10)

Table III - Mean and standard deviation values (in parentheses) of opening and closing gesture durations (ms), lip aperture (mm), peak velocities (mm/s), acceleration and deceleration durations (ms) in /pap/ syllables (token - PAPE)

		Pretonic	Tonic	Final(LOW)	Final(RIS)
		Opening gesture			
D.	AS	121(8)	136(13)	127(9)	107(4)
	BA	69(4)	76(5)	93(4)	114(5)
LA.	AS	9.6(.6)	9.9(.6)	9.6(.7)	8.7(.7)
	BA	7.8(.9)	10.3(.9)	10(1.2)	11(1.1)
Vp.	AS	155(22)	148(13)	161(8)	151(13)
	BA	173(26)	225(10)	187(19)	166(17)
Accel.	AS	62(9)	69(14)	69(8)	54(4)
	BA	39(3)	43(5)	60(7)	68(3)
Decel.	AS	59(7)	67(5)	58(7)	59(2)
	BA	30(1.5)	33(4)	32(3)	46(1.2)
		Closing gesture			
D.	AS	138(8)	163(11)	156(12)	180(10)
	BA	63(4)	71(3)	97(4)	108(3)
LA.	AS	11.9(1.5)	11(.8)	10.5(.9)	9.8(.6)
	BA	7.5(.8)	10.4(.8)	11.4(1.3)	11.8(1.2)
Vp.	AS	183(25)	161(13)	172(16)	132(12)
	BA	211(17)	267(29)	265(35)	225(35)
Accel.	AS	53(5)	59(3)	54(3)	64(3.5)
	BA	26(3)	29(2)	33(3)	46(4)
Decel.	AS	85(8)	105(11)	103(13)	116(12)
	BA	37(2)	42(3)	64(3)	61(2)

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 6.28, 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 /pa/ gestures although lip aperture is bigger and peak velocities higher in syllables with rising tone (F's 8.88, 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 generalizations 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 /pa/ syllables for speaker AS, and /pap/ syllables for BA. It can also be argued that the observed bigger apertures in word initial /pa/ syllables for AS are also an accent effect, given the increased predominance of word initial accent in spoken French. Speaker BA consistently accented the first syllable of "Papa" in sentences A (i) and (ii).

The localisation of the duration contrast to the tailend of closing /pap/ gestures suggests that pretonic closing gestures may be cut short by the opening gesture associated with the upcoming syllable in the sequence. In other words, gestural sliding, resulting in truncation of closing gestures may explain shorter gesture durations 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' /pa/ 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 /pa/ syllables without an accompanying duration difference suggest an articulatory manoeuvre not unlike that attributed to a stress contrast.

5. REFERENCES

[1] Delattre, P. (1966): A comparison of syllable-length conditioning among languages.*IRAL*,7,295-325.

[2] O'SHAUGHNESSY, D. (1981): A study of French vowel and consonant durations. *Journal of Phonetics*, 9,385-406.

[3] TOUATI,P. (1987): Structures prosodiques du suédois et du français. *Travaux de l'institut de linguistique de Lund*, 21.

[4] FLETCHER,J. AND VATIKIOTIS-BATESON,E. (1991): Prosody and articulatory timing in French. (Submitted)

[5] EDWARDS,J., BECKMAN, M.E., and FLETCHER, J. (1991): The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America*, **89**(1), 369-382.

[6] BENGUEREL, A-P. (1971): Duration of French Vowels in unemphatic stress. *Language and Speech*, 14, 383-391.

[7] Nelson, W.L. (1982): Physical principles for economies of skilled movements. *Biological Cybernetics*, 46, 135-147.

[8] Kirk, R.E. (1968): *Experimental design: Procedures for the behavioural sciences*. Belmont: CA: Wadsworth Publishing Co.

[9] SALTZMAN, E.L. AND MUNHALL, K.G.(1989): A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1(4), 333-382.

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 METHODE POUR LA RECHERCHE DE L'IMAGE CENTRALE : VOYELLES [i, e, a] DU FRANCAIS.

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 rythmical 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 METHODE

Notre étude porte sur la réalisation articulatoire des voyelles [i, e, a] du français sitées en fin de groupe rythmique ainsi qu'en milieu,en position interconsonantique (1 locuteur, grandeur réelle des mesures). Notre méthode d'analyse se fonde sur l'exploitation de films radiologiques avec synchronisation image/son (50 images par seconde) [1][2]. 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)

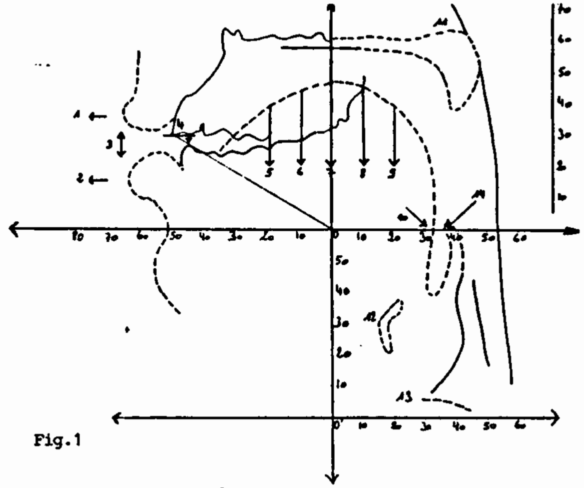
Dans notre corpus, nous relevons en fin de groupe rythmique : 17 ['i] , 6 ['e], 14 ['a] - en milieu de groupe rythmique : 7 [i], 1 [e], 10 [a]. Les voyelles sont précédées des consonnes suivantes : [p, b , ʃ, ʒ, s, k, g]. Il nous faudra tenir compte du contexte qui suit. Dans un premier temps nous nous sommes intéressée au comportement général de chacun des paramètres et nous avons chaque fois fait référence 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 degager 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 subit le moins l'influence du contexte, progressive et régressive. Etudions chaque paramètre :

Les lèvres : dans les deux cas, nous relevons une période de stabilité de trois images (14 à16). Nous savons que pour [s] les lèvres demeurent étirées comme pour [i]. En revanche, sous l'influence de la syllabe 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



Profil radiologique du conduit vocal.
Paramètres mesurés.

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).

Fig.2 (sibota'pi)

Images	14	15	16	17	18
Param.					
1	*67*	*67*	*67*	68	69
2	*66*	*66*	*66*	67	68
3	*11,5*	*11,5*	6	1	0
4	1	*1,5*	*1,5*	2	2
5	43	*45*	*45*	43	39
6	*48*	*48*	*48*	*48*	45
7	*48*	*48*	*48*	*48*	47
8	*43*	*43*	44	45	46
9	*35*	*35*	36	38	41
10	30	*28*	*28*	30	31
11	*68*	*68*	*68*	*68*	*68*
12	*(13;36')*	*(13;36')*	(13;35')	(13;34')	(13;33')
13	9'	*8'*	*8'*	*8'*	9'
14	35	*33*	*33*	*33*	34

Par 5,6,7 : Ces paramètres correspondent à la partie antérieure et centrale de la langue. Nous constatons que celle-ci s'élève dans les trois cas et 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).

Par 8 et 9 : Ces paramètres correspondent à la partie postérieure de la langue.Celle -ci subit très tôt l' influence de la voyelle postérieure [o]. En effet la langue va très rapidement s'élever. Nous optons pour la période où sa hauteur est la plus basse : Par 8 : 43 mm ; Par 9 : 35 mm (images 14 et 15) Les périodes choisies se placent au tout début de la voyelle.

Par 10 : Nous savons que pour [i] , la racine de la langue s'éloigne de la paroi pharyngale.

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 contxte 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 notamment 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 horizontal et vertical.

- 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.

La durée totale varie de 16cs à 22cs pour ['i]; 18cs à 22cs pour ['e]; 20cs à 26cs pour [a]. La durée s'allonge des voyelles fermées ['i] et ['e] à la voyelle ouverte

Fig.3

	['i]	['e]	['a]
Durée moyenne	9 images (18cs)	10 images (20cs)	11 images (22cs)
Image centrale	6ème	6ème ou 7ème	8ème

['a]. L'influence de certaines consonnes précédant les voyelles joue un rôle important quant à leur durée. Par exemple, [ʒ] et [s] réduisent la durée totale de ['i] à 16cs.

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 ['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 pour les voyelles 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

[1] BOTHOREL A.,SIMON P., WIOLAND F., ZERLING J. P.,1986, *Cinéradiographie des voyelles du français*, T.I.P.S.

[2] PASTOR A., 1988, Analyse inter-locuteurs [i] et [e] en français à partir de la radiocinématographie, Mémoire de D.E.A., Université Strasbourg II, 150 p.

[3] PASTOR A., 1989, "La notion de cible en termes de paramètres articulatoires", Séminaire S.F.A. & GRECO-PRC - CIRM, Marseille.

[4] SIMON P., 1967, *Les consonnes françaises, mouvements et positions articulatoires à la lumière de la radiocinématographie*, Klincksieck, Paris.

DE L'ANALYSE D'UNE VARIATION DE DEBIT DANS LA CHAINE PARLEE, A LA LUMIERE 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 case of French speech production. Presented in this study is an examination across rate conditions. 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

Ce travail fait partie d'une étude plus générale portant sur la réalisation articulatoire des consonnes occlusives sourdes et sonores, en position inaccentuée et à deux vitesses différentes de débit, sur la chaine parlée du français. Notre méthode d'analyse, la cinéradiographie, est fondée sur l'utilisation de documents expérimentaux associant les aspects articulatoires et acoustiques synchonisés (50 images / s., 2 locuteurs français, corpus identique enregistré à 2 débits). Ne seront traitées ici que les consonnes occlusives sourdes [p,t,k] à l'intervocalique (entourage identique) avec la consonne simple et double successive, pour 1 locuteur, grandeur réelle des mesures. Nous nous intéressons à la notion de débit comme variable articulatoire. Notre but à plus long terme est d'établir un classement des articulateurs en fonction de leur résistance à la variation du débit, et d'étudier la direction des faits de changement de débits en rapport avec les différents paramètres articulatoires. Les résultats pourraient contribuer à établir une stratégie des articulateurs en rapport avec le débit. S'il existe déjà de nombreux travaux sur l'articulatoire à un seul débit [1,2,10,12,16], le débit a dans la majorité des cas, intéressé des travaux à visée non articulatoire. Pour Miller [8], par la variation du débit, le nombre et la durée des pauses ainsi que la durée d'articulation sont modifiés. L'oreille compense certains phénomènes réductifs. Avec [9] elle a montré que les variations de la durée des pauses sont plus marquantes que celles de la vitesse d'articulation. Pour Malécot et coll. [7] il existe une corrélation positive entre longueur d'énoncé et débit syllabique (plus l'énoncé est long et plus le débit est élevé, et inversement). Pour Vaissière [13] il existerait une normalisation temporelle perceptive prenant en compte le débit de parole. Réduction et assimilation sont deux phénomènes observés sur le segment affecté d'un changement de débit. Mais au niveau suprasegmental, cela concerne essentiellement la représentation acoustique des mots. Wood [15] a remarqué une relative constance dans la durée syllabique malgré un changement de débit. Pour Shockey [11] il y aurait un lien causal entre débit et réduction phonologique. Gay et coll. [3,4,5] a montré avec la méthode E.M.G. qu'une augmentation de débit entraînait celle d'une activité musculaire. Avec la cinéradiographie il a mon-

tré un changement de cible sous l'effet d'une réduction. Gay [6] a étudié l'effet du débit sur la réalisation des cibles acoustiques des voyelles et la rapidité de mouvement pour les atteindre.

1.2 Corpus.

Phrases retenues et segments étudiés :
1. Il a pas mal. [ilapa'mal]
2. Les attabler. [lezata'ble]
3. Très acariâtre. [tʀezaka'ʀjatʀ]
4. Il zappe pas mal. [ilzappa'mal]
5. La chatte tachetée. [laʃattaʃ'te]
6. Trois sacs carrés. [tʀw asakka'ʀe]

1.3 Paramètres.

Nous avons relevé 14 paramètres (fig. 1)
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).
13 : base du larynx.
14 : épiglotte (mouvement horizontal et vertical).
Nous avons également relevé la distance d'occlusion de la langue ou les lèvres correspondant au lieu d'articulation des occlusives.

2. DUREES ET DEBITS.

2.1 Modalités d'enregistrement.

Il s'agit de phrases courtes chargées de sens en parole lue à deux débits différents: le lento et l'allegro. Chaque débit présente une régularité rythmique. La vitesse d'émission est le critère de variation d'un débit à l'autre. Nous entendons par lento une parole soutenue dans un style soigné, réalisé dans des conditions de lecture pour nous permettre d'obtenir des résultats comparables. L'allegro se différencie du lento par des caractéristiques appliquées souvent à la parole spontanée, c'est-à-dire la parole habituellement non lue et sans intention spéciale, si ce n'est une rapidité d'émission qui n'entrave pas pour autant la compréhension du message. Nous nous permettons cet abus de langage en appelant allegro un rythme de parole lue, si abus il y a [14].

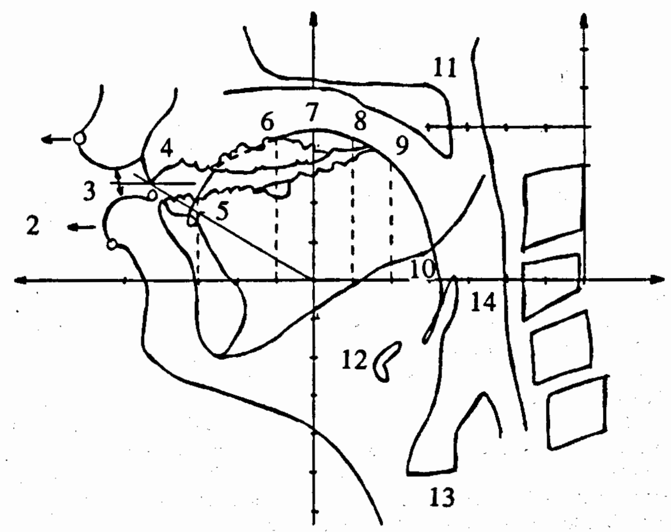


fig. 1 : profil radiologique du conduit vocal, paramètres mesurés.

2.2 Comportement temporel.

Les phrases du corpus (1.2) voient leur durée réduire entre le lento et l'allegro en moyenne de 30%. La durée des consonnes étudiées réduit en moyenne de 27% sauf pour l'extrabuccale double successive (ph. 4) pour laquelle 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 projetées 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'extrabuccale est réduite en allegro (de 20cs à10cs pour [pp]). La distance d'occlusion est supérieure en allegro (de 2mm). L'écartement bilabial est plus marqué en lento pour la vélaire.

Par. 4 : L'écartement du maxillaire est inférieur en allegro ([p] de 3mm, [pp] de 4mm, [k] et [kk] de 2mm).

Par. 5 : La distance d'occlusion pour l'alvéodentale est supérieure en allegro (de 2mm).

Par. 6 : partie de la langue plus élevée en lento (2mm).

Par. 7 : partie de la langue plus élevée en lento, surtout pour la vélaire (2mm) et l'alvéodentale (4mm).

Par. 8 : l'occlusion est retardée de 2cs en allegro pour la vélaire. Pour [kk] la durée d'occlusion est supérieure de 3cs en lento.

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 et 5mm. Les mesures sont décalées vers l'avant en allegro de 1mm.

Par. 11 : le sommet du voile est plus élevé en lento (sauf [t]) de 3mm pour [tt, pp] La distance d'occlusion est supérieure en allegro (sauf [p]) de 4mm pour [t], 7mm pour [tt], 6mm pour [k] et 3mm pour [kk]. Nous observons un creux dans la partie inférieure et postérieure du voile pour [p] 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 recule vers la paroi en moyenne de 6mm (3mm pour [k]) et se rapproche de la racine 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 articulatoire. Des différences de comportement des articulateurs peuvent être avancées à partir des variations des paramètres. Ainsi nous observons avec la perturbation apportée par le débit :

- Une réduction de l'amplitude du mouvement des articulateurs : par.1,2 et 3 lèvres moins avancées avec un écartrment moins marqué ; par. 4 écartement du maxillaire réduit ; par. 6,7,8 et 10 langue moins élevée et mouvement de la racine réduit ; par. 11 sommet du voile moins élevé et distance d'occlusion avec la paroi pharyngale supérieure. Quand la partie postérieure du voile présente un creux son écartement avec la paroi est réduit en allegro.
- Un décalage du mouvement des articulateurs dans l'espace : par. 10 et 14 mouvements de la racine et de l'épiglotte décalés vers la paroi ou vers l'avant ; par. 11 le creux du voile est situé plus haut en allegro, sinon le voile se rapproche de la paroi ; par. 12 l'os hyoïde a tendance à être à la fois plus bas et plus reculé.

4. CONCLUSION.

La chaîne parlée observée à travers une 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 articulatoire touchant plus ou moins certains articulateurs, et provoquant un phénomène de compensation interarticulateurs.

5. REFERENCES.

[1] BOTHOREL A.,SIMON P., WIOLAND F., ZERLING J. P.,1986, *Cinéradiographie des voyelles du français, Recueil de documents synchronisés pour quatre sujets : vues latérales du conduit vocal, vues frontales de l'orifice labial, données acoustiques*, T.I.P.S.

[2] BRICHLER-LABAEYE C., 1970, *Les voyelles françaises, mouvements et positions articulatoires à la lumière de la radiocinématographie*, Klincksieck, Paris.

[3] GAY T., HIROSE H., 1973, "Effect on Speaking Rate on Labial Consonant Production", *Phonetica*, 27, (44-56).

[4] GAY T., USHIJIMA T., HIROSE H., & COOPER F.S., 1974, "Effect on Speaking Rate on Labial Consonant-Vowel Articulation", *Journal of Phonetics*, 2, (47-63).

[5] GAY T., USHIJIMA T., 1974, "Effect on Speaking Rate on Stop Consonant-vowel Articulation", *Speech Comm. Seminar Stockholm*, 1-3, (205-208).

[6] GAY T., 1977, "Effect on Speaking Rate on Vowel Formant Movments", *Status Report on Speech Research*, Haskins Lab., (101-117).

[7] MALECOT A., JOHNSTON R., & KIZZIAR P.A., 1972, "Syllabic Rate and Utterance Length in French", *Phonetica*, 26, (235-251).

[8] MILLER J.L., 1981, "Effects on Speaking Rate on Segmental Distinctions", *Perspectives on the Study of Speech*, Eimas & Miller, (39-74).

[9] MILLER J.L., GROSJEAN F., & LOMANTO C., 1984, "Articulation Rate and its Variability in Spontaneous Speech", *Phonetica*, 41, (215-225).

[10] ROCHETTE C.E., 1973, *Les groupes de consonnes en français, étude de l'enchaînement articulatoire à l'aide de la radiocinématographie et de l'oscillographie*, Klincksieck, Paris.

[11] SHOCKEY L., 1987, "Rate and Reduction : Some Preliminary Evidence", *In Honor of Ilse Lehiste*, Channon & Shockey, (217-224).

[12] SIMON P., 1967, *Les consonnes françaises, mouvements et positions articulatoires à la lumière de la radiocinématographie*, Klincksieck, Paris.

[13] VAISSIERE J., 1989, Thèse d'habilitation, Strasbourg.

[14] VAXELAIRE B., 1989, "Etude comparée de la durée des consonnes en français dans des séquences VCV, lues à deux vitesses de débit", Mémoire de D.E.A., préparé sous la direction de M. le Professeur A. Bothorel, Strasbourg.

[15] WOOD S., 1973, "What Happens to Vowels and Consonants when we Speak Faster ?", *Working Papers*, Phonetics Lab., Lund University, 9, (8-39).

[16] ZERLING J.P., 1990, *Aspects articulatoires de la labialité vocalique en français. Contribution à la modélisation à partir de labiophotographies, de labiofilms et flims radiologiques.*, Thèse d'Etat, Strasbourg.

Nos sincères remerciements à Pierre-Yves Connan qui nous a aidée 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 examined by comparative analysis of over 1,500 elicited 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 events 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 error behavior (reviewed in: Cutler, 1982). Speech errors of aphasics have also demonstrated the same, consistent pattern (Blumstein 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 the phonotactic constraints of the first language.

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

2. PROCEDURE

A speech-error elicitation task, modelled on one created by Shattuck-Hufnagel (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.

Forty 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 CVCVC 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: parade fad foot parole;
French: parade fad foot parole.
The remaining five word sets were different in the two languages. These did not all have the same syllable structure. All sets included segments unique to each language in word-onset position. Example: (Target segment: TH)
English: six <u>th</u>ick <u>th</u>istle 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 uttered segment involved in an error occur in the word string, the direction of the error (either anticipatory or perseveratory) and the relative position in word of the target and the uttered segment in the speech error were recorded. Stress was also noted, for both the target and uttered segments, as well as voicing and place of articulation.

For intrusion errors, in which the uttered segment in an error does not occur in the stimulus set, comparison was made between the target segment and the uttered segment for syllable structure, place of articulation, and rhyme. The number of segments replaced 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, calucations 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 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 position to those in diferent word position (initial/initial to initial/medial, etc.); (3) stress effect -- the ratio of interacting segments bearing similiar lexical stress to those bearing different lexical stress; (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 patterns 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 60%) in word-final position. Two 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 in both first and second languages of bilinguals.

Several characteristics of errors produced by bilinguals in their first and second languages were significantly different from those of the monolingual control groups. These differences included: error position, size and type.

Error position.

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 English. 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 single segments only. The only errors of bilinguals which involved units greater than a single segment were "blend" errors, a combination of syllables from two words in the stimulus set, in the first language.

Error type.

a. Blend. Although "blend" errors were made by almost all monolingual speakers, very few blends were made by bilinguals, and all in their first languages. No "blend" errors occured in the second language of bilinguals. All L2 errors were restricted to single segments.

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 monolinguals ranged from 1-5 segments in size. Bilingual 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).

Rhyme. 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.

No 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 occured 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, 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 in French/English bilinguals affects the phonological organization of speech production 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 "mobility" of a segment, its occurence as a substitution for another segment in positions or words other than its target position, is considered evidence of "independent" behavior, it might be concluded that L2-only segments do not function independently. The 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 particpation in the processing of second language speech of bilinguals, which may involve a more holistic functions (reviewed in Fabbro et al. 1990). Further study of other bilingual populations is indicated to further explore the "universality" issue, and the effects of bilingualism.

REFERENCES.

Blumstein, S.(1990). Phonological Deficits in Aphasia, Cognitive Neuropsychology and Neurolinguistics, A. Caramazza, (ed.), Lawrence Erlbaum, Hillsdale, NJ.

Cutler, A. (1982). The Reliability of Speech Error Data, Slips of the Tongue and Language Production, A. Cutler (ed.), New York: Mouton, 7-29.

Fabbro, F., Gran, L., Basso, G., Bava, A. (1990). Cerebral Lateralization in Simultaneous Interpretation, Brain and Language, 39, 69-89.

Fowler, C. (1987). Current Perspectives on Language and Speech Production: A Critical Overview, Language and Speech Production, 194-277.

Shattuck-Hufnagel, S. (1987). Word-Onset Consonants in Speech Production Planning, 114th Meeting of the Acoustical Society of America.

Vaid, J. (1986). (Ed.) Language Processing in Bilinguals: Psycholinguistic and Neuropsychological Perspectives, Lawrence Erlbaum Assoc., Inc., Hillsdale, NJ.

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

V. Berthier, C. Abry, T. Lallouache

Institut de la Communication Parlée, 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.

L'une des phases de la fabrication traditionnelle des sifflets d'écorce, au printemps, met en jeu une telle coordination. C'est après avoir effectué plusieurs opérations préparatoires, que le sujet saisit son couteau par la lame – entre le pouce et la phalangine de l'index, dans une *prise par opposition pulpo-latérale* [2] – pour battre en rythme un petit tronçon de frêne. Ce faisant, il scande en dialecte une formulette d'incantation à la sève (cf. *Annexe*). Cette étape a pour but de *réussir* à détacher du bois le manchon d'écorce.

Ce geste du bras produit, en l'occurrence, une séquence de *percussions perpendiculaires lancées diffuses* [5]: nous l'appellerons "geste de volée", comme la *volée* du marteau.

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

2. CINÉMATIQUE DU GESTE DE VOLÉE

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 frêne, 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 en vue latérale. 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-doigt;
(c) Intersection lame-virole du couteau;
(d) Moëlle du rameau de frêne.

Un poste de numérisation et de traitement d'images [4] nous a permis de mesurer différents paramètres kinésiologiques, 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* virole-moëlle. Ces paramètres, édités en fonction du temps (Fig. 2), ont rendu possible le repérage de plusieurs relations de phasage: la *distance* diminue à mesure que la valeur de l'*angle* augmente; et elle atteint son minimun 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 volée (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);
-*percuté* (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 46 % du cycle. Les études de gestes traditionnels comparables sont rares. Une recherche ethnotechnologique, réalisée en Normandie [1], nous a permis cependant de comparer diverses percussions, dont celle d'un bourrelier, qui assouplit le cuir avec son marteau rivoir. Avec une durée moyenne de cycle de 234 ms, décomposable en trois phases – une descente (*lancé*), un contact (qui correspond à notre phase de *percuté*) 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 efficace de *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 SIGNAL DE PAROLE

Sur le signal audio, échantillonné à 16 KHz, la mesure du cycle de percussion se précise, confirmant sa régularité : la variation *maximale* autour 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 récitation de la formulette.

L'étude de la relation temporelle entre le pic de percussion et le début de la voyelle suivante (c'est-à-dire l'établissement d'une structure formantique définie) a fait apparaître une variation importante, de 0 à 100 ms. Lorsqu'on examine la distribution de ces percussions, on constate pourtant que celles-ci ne se produisent jamais avant la fin des voyelles précédentes. 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 démarrage des séquences, le geste se cale d'abord sur la parole: ce que révèlent certains coups donnés "à vide". 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 voyelles, 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 permet pourtant de constater que la fréquence d'ouverture et de fermeture du tractus vocal – qui correspond au rythme syllabique régulé par la mandibule (soit 6Hz ± 1 [9]) –, peut aisément s'ajuster à la fréquence des battements régulés par le couple main-bras (qui est de 4-6 Hz en *cadence rapide* [7]). Cette coordination rythmique du geste et de la parole semble donc ici ralentir la fréquence de modulation du conduit vocal, puisque celle-ci est *entraînée* 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 du Musée Dauphinois et du PPSH Rhône-Alpes n°20.

5.RÉFÉRENCES

[1] BRIL, B. (1986), *Appartenance régionale et identité culturelle*, Rapport mission du Patrimoine.

[2] KAPANDJI, A. (1980), *Physiologie articulaire, membre supérieur,* Paris : Maloine S.A.

[3] KLAPP, S.T. (1979), Doing two things at once: the role of temporal compatibility, *Memory & Cognition, 7 (5)*, 375-381.

[4] LALLOUACHE, M.T. (1990), "Un poste "visage-parole". Acquisition et traitement de contours labiaux", *18° J.E.P., Montréal*, 282-291.

[5] LEROI-GOURHAN, A. (1943), *L'homme et la matière,* Paris : Albin Michel.

[6] MARCUS, S. (1975), *Perceptual centers,* Unpublished fellowship dissertation, Cambridge, King College.

[7] NEILSON, P.D. (1972), *Med. and Biol. Eng., 10,* 450-459.

[8] SCHMIDT, H.G. (1969), *Les postures de travail défavorables,* C.E.E.

[9] SOROKIN, V.N., GAY, Th. & EWAN, W.G. (1980), "Some biomechanical correlates of the jaw movements", *J. Acoust. Soc. Am.,.Suppl.1, Vol. 68,* S32.

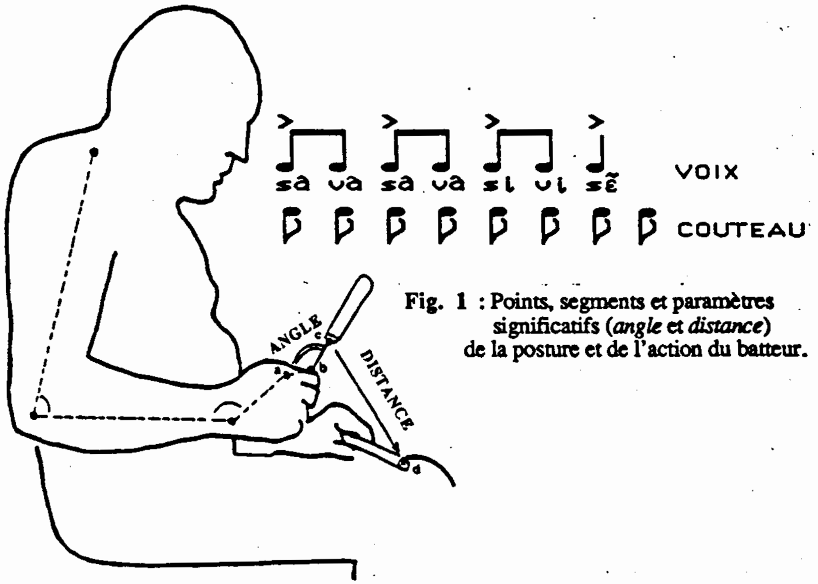


Fig. 1 : Points, segments et paramètres significatifs (*angle* et *distance*) de la posture et de l'action du batteur.

ANNEXE

['sava 'sava si vi'sẽ	Sève! Sève! Saint Vincent!
meta 'pajɔ meta 'fẽ	Moitié paille, moitié foin,
sava'reta pa si 'bjẽ	"Sèverette" pas si bien
kɔ la 'mɛrda do po'jẽ	Que la merde du poulain
vẽ 'bjẽ	Viens bien!
di 'rẽ	Dis rien!
eklapa 'rẽ]	Fends "rien"!

(P.M., 65 ans, Autrans, 12-5-88; formulette du sifflet, en dialecte francoprovençal)

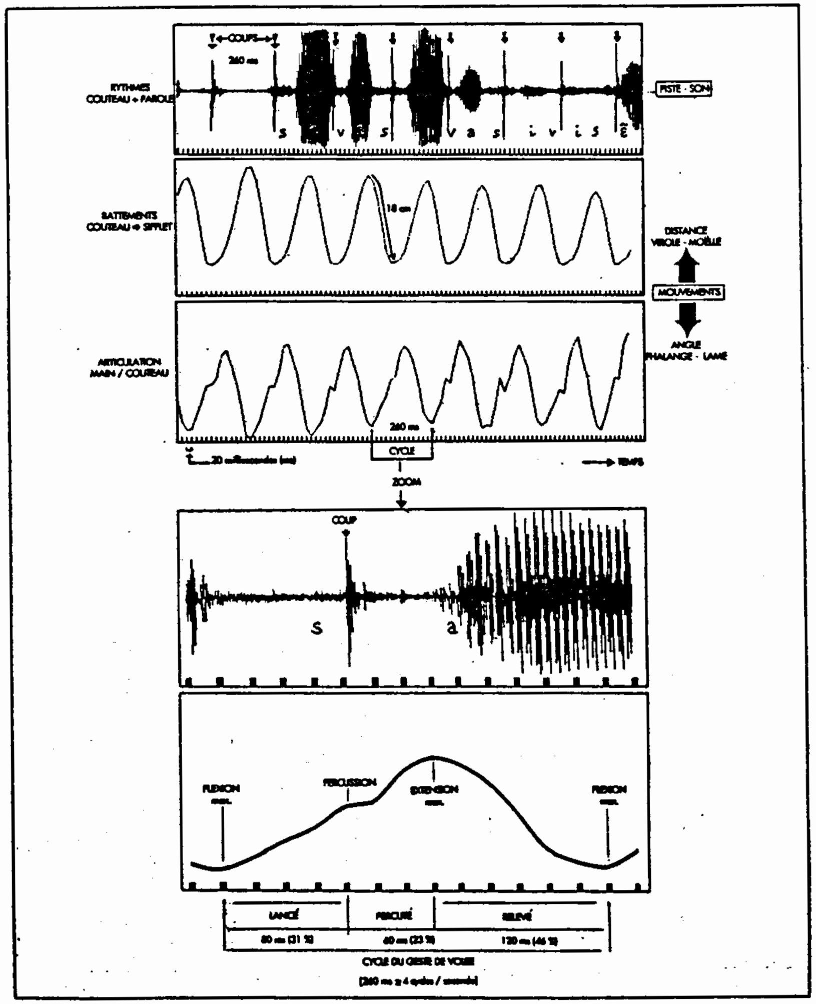


Fig. 2 : Signal de parole et cinématique du geste de volée.

INTEGRATION OF AUDITORY AND VISUAL COMPONENTS OF ARTICULATORY INFORMATION IN THE HUMAN BRAIN

Reijo 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 "McGurk 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 videotaped face articulating either [pa] or [ka]. The same auditory 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 by visual cues alone [4].

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 "McGurk 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 [ga] eliciting an auditory perception of [da] [8, 10]. 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 proposed, on the basis of brain damages, that the ability to lip read is a function 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 where they were discordant (V≠A, 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≠A 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 neuromagnetic 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 "silent identification". 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 stimulus 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 samples 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 difference 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≠A stimuli either as [ta] or [ka] or something in between.

The magnetic responses to the frequent V=A stimuli typically consisted of three consecutive deflections, peaking at 50, 100, and 200 ms (Fig. 1). Similar deflections are elicited by any kind of abrupt sounds and can be explained by equivalent current dipoles in the supratemporal auditory cortex.

The magnetic responses to infrequent V≠A 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 infrequent ones) was elicited by infrequent V=A stimuli among frequent V≠A stimuli. However, the signals to the auditory syllables preceded by frequent green and infrequent red light stimuli were identical (Fig. 1).

The infrequent V≠A stimuli elicited a distinct difference waveform in 7 out of the 10 subjects. Infrequent V=A stimuli elicited such a waveform in 6 out of 7 subjects studied, including those three who did not show it to infrequent V≠A 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

[1] BINNIE, C.A., MONTGOMERY, A. A. & JACKSON, P.L. (1974), "Auditory and visual contributions to the perception of consonants", *Journal of Speech and Hearing Research*, 17, 619-630.

[2] CAMPBELL, R. (1987), "The cerebral lateralization of lip-reading". In *Hearing by eye: The psychology of lip-reading* (B. Dodd & R. Campbell, eds), 215-226. London: Lawrence Erlbaum.

[3] DODD, B. (1977), "The role of vision in the perception of speech", *Perception*, 6, 31-40.

[4] FISHER, C.G. (1969), "The visibility of terminal pitch contour", *Journal of Speech and Hearing Research*, 12, 379-382.

[5] GESCHWIND, N. (1965), "Disconnexion syndromes in animals and man", *Brain*, 88, 237-294 & 585-644.

[6] HARI, R. & LOUNASMAA, O.V. (1989), "Recording and interpretation of cerebral magnetic fields", *Science*, 244, 432-436.

[7] JACKSON, P.L. (1988), "The theoretical minimal unit for visual speech perception: Visemes and coarticulation", *The Volta Review*, 90, 99-115.

[8] MACDONALD, J. & MCGURK, H. (1978), "Visual influences on speech perception processes", *Perception & Psychophysics*, 24, 253-257.

[9] MASSARO, D.W. & COHEN, M.M. (1983), "Evaluation and integration of visual and auditory information in speech perception", *Journal of Experimental Psychology: Human Perception and Performance*, 9, 753-771.

[10] MCGURK, H. & MACDONALD, J. (1976), "Hearing lips and seeing voices", *Nature*, 264, 746-748.

[11] MILLS, A.E. & THIEM, R. (1980), "Auditory-visual fusions and illusions in speech perception", *Linguistische Berichte*, 68/80, 85-108.

[12] NÄÄTÄNEN, R., GAILLARD, A. W.K. & MÄNTYSALO, S. (1978), "Early selective attention effect reinterpreted", *Acta Psychologica*, 42, 313-329.

[13] SAMS, M., AULANKO, R., HÄMÄLÄINEN, M., HARI, R., LOUNASMAA, O.V., LU, S.-T. & SIMOLA, J. (in press), "Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex", *Neuroscience Letters*.

[14] SAMS, M., HÄMÄLÄINEN, M., ANTERVO, A., KAUKORANTA, E., REINIKAINEN, K. & HARI, R. (1985), "Cerebral neuromagnetic responses evoked by short auditory stimuli", *Electroencephalography and Clinical Neurophysiology*, 61, 254-266.

[15] SUMMERFIELD, Q. (1987), "Some preliminaries to a comprehensive account of audio-visual speech perception". In *Hearing by eye: The psychology of lip-reading* (B. Dodd & R. Campbell, eds), 3-51. London: Lawrence Erlbaum.

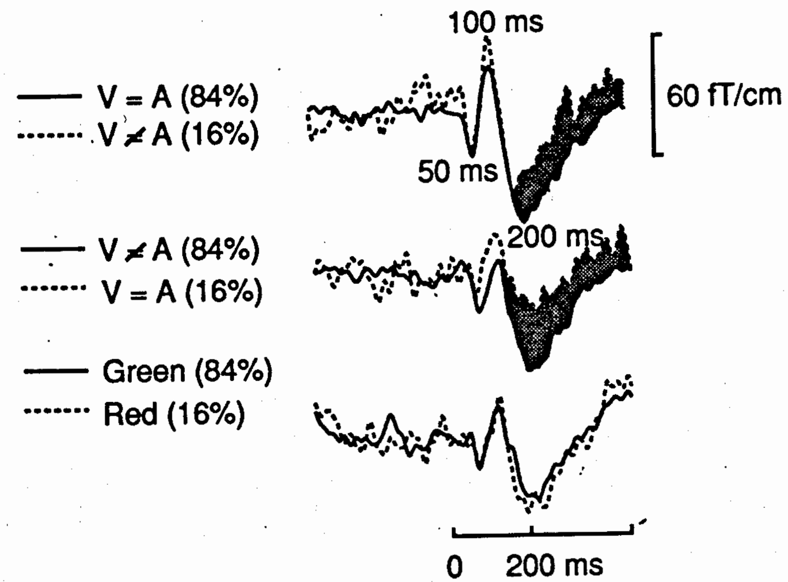


FIGURE 1. Magnetic responses of one subject, measured with a 24-SQUID gradiometer over the left hemisphere in three measurement conditions. Only one of the channels with the largest responses is shown. The three pairs of traces were recorded over the same area in consecutive measurements. The number 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 40 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 experimentor ; 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-phonatory 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 front ways of talking 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 experimentor. 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 tourbilllons! Les tourbillons étaient trop forts!"

The selected syllable was [jɔ̃] in "tourbillons". The results of this analysis [4] 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.

- [bijɔ̃]

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

- [ə]

We noticed that the following cue :

[F4 - F3]

could be dependent of speaker's vocal behaviour : when converting these frequential 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 [jɔ̃] in "tourbillons". We selected the slope of F2 between [j] and [ɔ̃] 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 unsteady 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 differenciation 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

[1] BOOMER D.S., LAVER J.D.M. (1968) : Slips of the tongue. *British Journal of Disorders of Communication*, 3, 1-11.
[2] HARDCSTLE W.J. (1976) : *Physiology of Speech Production*, Academic Press, London.
[3] LENNEBERG E.H. (1967) : *Biological Foundations of Language*. New York : John Wiley and sons.
[4] LHOTE E., ABOU HAIDAR L. (1990) : Speaker verification by a vocal proxemy cue. *ESCA Tutorial and Research Workshop on Speaker Characterisation in Speech Technology*, Edinburgh, 149-154.
[5] LIENARD J.-S. (1989) : Variabilité, contrainte et spécificité de la parole : un cadre théorique. Actes du Séminaire : *Variabilité et Spécificité du locuteur*, CIRM - Marseille, Luminy.
[6] MAC NEILAGE P.F., DE CLERK J.L. (1969) : On the motor control of coarticulation in CVC mono-syllables. *Journal of the Acoustical Society of America*, 45, 1217-1233.
[7] OHMAN S.E.G. (1966) : Co-articulation in VCV utterances : spectrographic measurements. *Journal of the Acoustical Society of America*, 39, 151-168.
[8] PASCAL D. (1989) : Etude de la similarité des voix masculines : corrélation entre mesures physiques et structure perceptive. Actes du Séminaire : *Variabilité et Spécificité du locuteur*, CIRM - Marseille, Luminy.

6. FIGURES

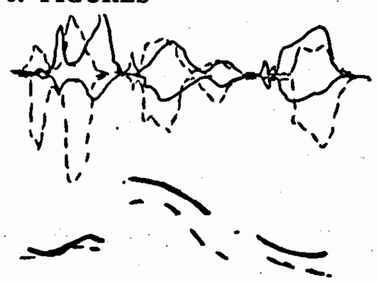


FIGURE 1 - Two Speakers close at the rhythmic and melodic level
Sentence : "C'est d'accord ou quoi ?"

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

FIGURE 4 - Formant trajectories of two distant speakers

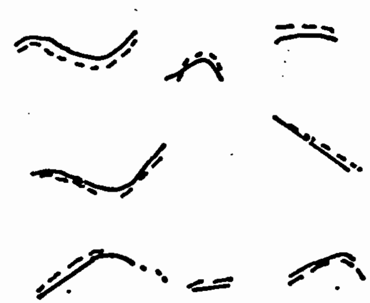


FIGURE 3 - Formant trajectories of two close speakers

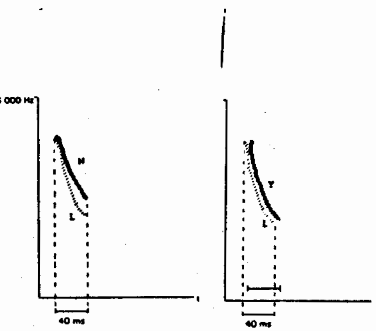


FIGURE 5 - F2 transition slope in the syllable [ijɔ̃]

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 step 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 17.5dB 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 they 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 broad-band 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). Filter 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 dBA SPL in broad-band 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 70.7% correct 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 noise-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 1: mean DLs and standard errors (Hz) for experiment 1

modul:	none	5 Hz	10 Hz
250 ms			
1 peak	30.29	38.27	37.64
sem	*1.64*	*2.05*	*1.89*
2 peak	44.59	37.55	39.34
sem	*0.99*	*2.28*	*2.09*
500 ms			
1 peak	25.58	33.83	33.41
sem	*2.26*	*2.20*	*1.28*
2 peak	41.34	32.18	35.22
sem	*2.19*	*2.57*	*1.48*

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 centre 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 endows 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 2: mean DLs and standard errors (Hz) for experiment 2 (Target Peak frequency 1500 Hz).

modul:	none	5 Hz	10 Hz
250 ms			
1 peak	37.26	46.12	39.60
sem	*3.80*	*2.19*	*3.29*
2 peak	50.88	42.29	44.24
sem	*4.69*	*2.51*	*3.92*
500 ms			
1 peak	30.90	36.27	34.97
sem	*2.86*	*3.62*	*3.17*
2 peak	45.68	37.95	35.66
sem	*4.19*	*2.74*	*3.60*

TABLE 3: mean DLs and standard errors (Hz) for experiment 2 (Target Peak frequency 900 Hz).

modul:	none	5 Hz	10 Hz
250 ms			
1 peak	32.52	28.60	22.68
sem	*4.61*	*6.39*	*5.57*
2 peak	38.45	29.95	21.45
sem	*5.13*	*5.65*	*4.71*
500 ms			
1 peak	31.25	21.51	21.28
sem	*4.00*	*5.54*	*4.41*
2 peak	37.23	25.04	23.42
sem	*4.24*	*5.14*	*5.49*

The pattern 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 discrimination modulation affords derives from properties 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

[1] McAdams S.(1989) J.Acoust. Soc.Amer., 86, 2148-2159
[2] Rasch R.(1978) Acustica, 40,21-33

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

M.-A. Cathiard *, G. Tiberghien *, A. Cirot-Tseva **, M.-T. Lallouache **, P. Escudier **

* 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; (iiii) 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 [zizy] 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 prosodic 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.*

Fig.1

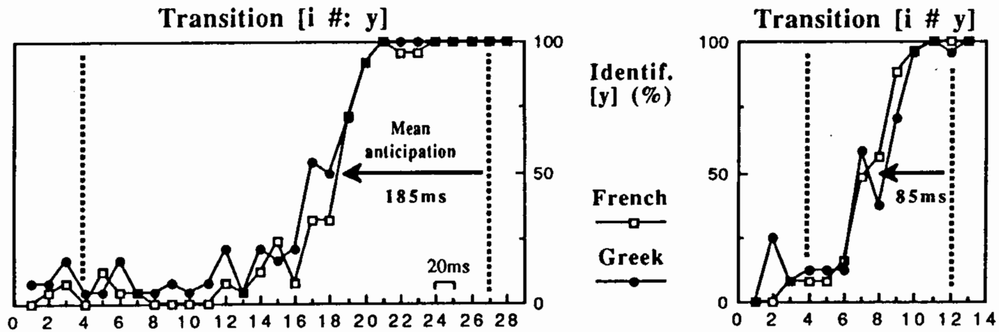


Fig.2

Figures 1-4. – Above : identification functions of [i -> y] transitions for 25 French and 24 Greek subjects.
– 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].

Fig.3

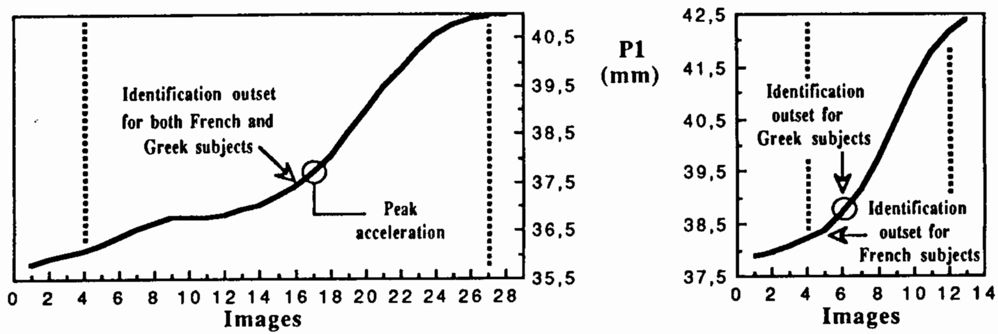


Fig.4

We will try to test these models by analysing *articulatory and visual* data.

2.METHOD

2.1.Corpus

We used [i # y] transitions which were embedded in a carrier sentence : "tu dis : UHI ise?" [t y d i # y i i: z], "you say : ...", where UHI is, by convention, an "Indian name" and "ise" a third person present of a nonsense verb "iser". [t y d i # i i i: z] is the control stimulus with IHI 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 recorded in random order.

2.2.Video recording

A French male talker was filmed, at 50 frames/second, with simultaneous face and profile views, in a sound-proof booth. Talker's lips were made-up in blue : a Chroma-key was connected to the output of the front camera so that the blue color was transformed to saturated black in real time in order to realize a maximal outlines detection of the lip slit. The subject wore black sunlight goggles in order to protect his eyes against the 1000 W halogen floodlight; a slide rule was fixed on the right side of the goggles to ensure adequate profile articulatory measurements [4].

2.3.Selection of visual stimuli

2.3.1.Acoustic measurements

Four utterances were selected among 40 after duration measurements of all intervocalic pauses. They were chosen as representative of mean durations for the short pause (# = 160 ms) and the long one (#: = 460 ms).

2.3.2.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 cubic 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, occuring 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 occured 260 ms before the acoustic onset of the [y] in [i # y] versus 560 ms in [i #: y] (i.e. the protrusion gesture began 100 ms into the [i] vowel). Finally our articulatory stimuli correspond better to a LA model, with respect to *dates* of onsets, but they display rather H *profiles* (fig. 3 & 4).

2.4.Test procedure

We selected 13 images for short transitions and 28 images for the long ones, with 3 images before pause onset and 1 after pause termination. We thus obtained a total of 82 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 subjects with the task. The stimuli were displayed individually to each subject on a high resolution computer screen. The task was to decide whether the speaker was uttering [i] or [y]. Subjects were encouraged to answer rapidly (within a few seconds) via a computer mouse.

2.5.Subjects

25 French and 24 Greek normal-hearing native speakers served as naive subjects (their hearing and vision acuities were checked). A good auditory 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 control transitions displayed steady state profiles, since [i –> i] images were generally identified as [i] (above 80%). Subjects were able to identify correctly (at 100%) "targets" of the presented vowels, i.e. images corresponding to the non silent outsets of [y]. Moreover, they were clearly able to capture anticipated segmental information on rounding (95% correct) *up to 120 ms before the acoustic onset of the vowel, be they 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, thus delivering information on the possible similarity in *steepness* between the boundaries.

For [i # y] : 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], within 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 protrusion, on the other hand, the temporal *accuracy* of these boundaries (i.e. their *steepness* estimated by functions gradients) did not depend on anticipation : *80 to 110 ms 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 toward [i]). The other way round they could have read the "U" choice as [u]. In both cases ([y] or [u]) however, they did not capture significantly less *rounding* anticipation than French did.

3.2.Visual perception of anticipation and articulatory models.

The observed significant shifts in boundaries could by themselves discredit the prediction of a time-locked visual anticipation. In fact, our perceptual as our articulatory (cf. 2.3.2.) data allow 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 until it is clearly *accelerated* (fig. 3 & 4). More precisely, it is the position of the visual identification *outset* (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 [20 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 perceivable 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 LA, TL nor H.

The cross-language comparison did not revealed significant discrepancies in visuo-temporal 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 quest.

* Many thanks to J.-L. Schwartz and W. Serniclaes for their advices in Probit Analysis and to T. Brennen for improving our English.

5.REFERENCES

[1] BENGUÉREL, A.P. & COWAN, H.A. (1974), "Coarticulation of upper lip protrusion in French", *Phonetica, 30*, 41-55.

[2] ESCUDIER, P., BENOIT, C. & LALLOUACHE M.-T. (1990), "Visual perception of anticipatory rounding gestures", *JASA, Suppl. 1, 87*, S126.

[3] FINNEY, D.J. (1971), *Probit analysis*, Cambridge University Press.

[4] LALLOUACHE, M.-T. (1990), "Un poste 'visage-parole'. Acquisition et traitement de contours labiaux", *Actes des XVIIIèmes J.E.P. (28-31 Mai)*, Montréal, Canada, pp. 282-291.

[5] McGURK, H. (1981), "Listening with eye and ear" (paper discussion), in T. Myers, J. Laver & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 394-397), Amsterdam, North-Holland.

[6] MASSARO D.W. (1987), *Speech perception by ear and eye : a paradigm for psychological inquiry*, Lawrence Erlbaum Associates.

[7] PERKELL, J.S. (1990), "Testing theories of speech production : implications of some detailed analyses of variable articulatory data", in W.J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling*, (pp. 263-288), Kluwer Academic Publishers, Dordrecht, Boston, London.

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 murmur 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, Grose et al. 1989). Several studies indicate that hearing-impaired persons show deteriration of temporal resolution (Fitzgibbons and Wightman 1982, Fitzgibbons and Gordon-Salant 1987, Glasberg et al. 1987, Nelson and Freyman 1987, Moore and Glasberg 1988, Grose 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 altered only (Parady 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.

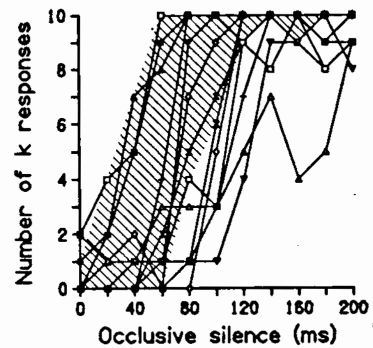
Samples of natural speech tokens "aka" and "aga" were recorded from an adult male speaker. Speech waveforms were edited in a computer. From the "aka" sound eleven tokens were formed by varying occlusive silence duration from 0 to 200 milliseconds in 20 milliseconds steps. From the "aga" sound one cycle of waveform during the voiced murmur was selected as having the same fundamental frequency as the "aka" sound. Bursts of murmur were then constituted by concatenations of this cycle and multiplication by a trapezoidal 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 adding these various bursts at the end of the first "a" of the "aka" sound thus constituting "a+voiced murmur" 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 succesively. The first was always the first "a" of the "aka" item and the second was one of the various "a+voiced murmur" token. Each "a+voiced murmur" was presented ten times randomly and the subject was asked to indicate whether the stimuli 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 normals are presented 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 shft along the time axis keeping a steepness similar to normal curves.

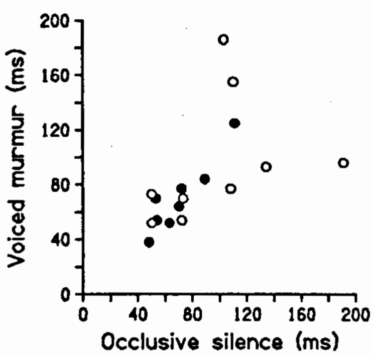
Figure 1



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 indicate the high value of 0.925 significant at the 0.001 level. The results from hearing-impaired ears show a considerable scatter, associated with a correlation coefficient of 0.354 not statistically significant. Hearing-impaired ears with normal results at experiment 1 also showed normal results at experiment 2, only results outside normality show a considerable scatter. If all normal results are considered whether coming from normal or pathological ears ; a correlation of 0.836 significant at the 0.01 level. In the group of pathological ears correlations were further considered with the following audiological data : etiology of deafness, age of the patient, duration of deafness, and auditory thresholds at octave frequencies from 250 to 8000 Hz. Only two correlations were found significant at the level of 0.05. They linked auditory thresholds at 250 and 500 Hz with results of experiment 1. It must be noted however that the two worst scores at experiments 1 and 2 were observed for patients diagnosed as probable Ménière.

Figure 2



Duration of occlusive silence versus duration of voiced murmur, both giving a score of 75% success for all subjects of these experiments. Black dots : normal ears, circles : pathological ears.

4 - DISCUSSION

Results of this study show an abnormally 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 originate from a deteriorated temporal resolution at voicing frequency.

The observed temporal shift in the hearing-impaired indicate 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 (Picheny 1986, 1989). It also indicate 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 impairement 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 (Cazals 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 Cnamts-Inserm.

5 - REFERENCES

Cazals Y and Horner K Abnormal two-sound interactions in hydropic cochleas of the guinea pig. In Basic Issues in Hearing, Duifhuis, Horst and Wit eds, Academic press, 1988, p 457-465.

Dorman MF, Marton K, Hannley MT, Lindholm JM (1985) Phonetic identification by elderly normal and hearing impaired listeners. J. Acoust. Soc. Am., 77, 664-670.

Fitzgibbons PJ and Gordon-Salant S (1987) Minimum stimulus levels for temporal gap resolution in listeners with sensorineural hearing loss. 81, 1542-1545.

Fitzgibbons PJ and Wightman FL (1982) Gap detection in normal and hearing-impaired listeners. J. Acoust. Soc. Am., 72, 7671-765.

Ginzel A., Brahe Pedersen C, Spliid PE and Andersen E (1982) The role of temporal factors in auditory perception of consonants and vowels. Scand. Audiol. 11, 93-100.

Glasberg BR, Moore BCJ, Bacon SP (1987) Gap detection and masking in hearing impaired and normal hearing subjects. J. Acoust. Soc. Am., 81, 1546-1556.

Green DM and Forrest TG (1989) Temporal gaps in noise and sinusoids. J. Acoust. Soc. Am., 86, 961-970.

Grose JH, Eddins DA and Hall III JW (1989) Gap detection as a function of stimulus bandwidth with fiwed high-frequency cutoff in normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am., 86, 1747-1755.

Johnson D, Whaley P and Dorman MF (1984) Processing of cues for stop-consonant voicing by yound hearing-impaired listeners. J. Speech and Hearing Res., 27, 112-118.

Lisker L (1957) Closure duration and the intervocalic voiced distinction in english. Language, 33, 42-49.

Lisker L (1981) On generalizing the rabid-rapid distinction based on silent gap duration. Haskins Lab. Status report on speech research SR-65, 251-259.

Moore BCJ and Glasberg BR (1988) Gap detection with sinusoids and noise in normal, impaired and electrically stimulated ears. J. Acoust. Soc. Am., 83, 1093-1101.

Nelson DA, Freyman RL (1987) Temporal resolution in sensorineural hearing impaired listeners. J. Acoust. Soc. Am;, 81, 709-720.

Parady S, Dorman MF and Whaley P (1981) Identification and discrimination of a synthesized voicing contrast by normal and sensorineural hearing-impaired children. J. Acoust. Soc. Am., 69, 783-790.

Picheny MA, Durlach NI and Braida LD (1986) Speaking clearly for the hard of hearing II : Acoustic characteristics of clear and conversational speech. J. Speech and Hearing Res., 29, 434-446.

Picheny MA, Durlach NI and Braida LD (1989) Speaking clearly for the hard of hearing III : An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech. J. Speech and Hearing Res., 32, 600-603.

Port RF (1981) Linguistic timing factors in combination. J. Acoust. Soc. Am., 69, 262-274.

Price PJ and Simon HJ (1984) Perception of temporal differences in speech by "normal hearing" adults : effects of age and intensity. J. Acoust. Soc. Am., 76, 405-410.

Revoile S, Pickett JM and Holden LD (1982) Acoustic cues to final stop voicing for impaired- and normal-hearing listeners. J. Acoust. Soc. Am., 72, 1145-1154.

Shailer MJ and Moore BCJ (1983) Gap detection as a function of frequency, bandwidth and level. J. Acoust. Soc. Am., 72, 467-473.

Shailer MJ and Moore BCJ (1985) Detection of temporal gaps in bandlimited noise : Effects of variations in bandwidth and signal-to-masker ratio. J. Acoust. Soc. Am., 77, 635-639.

Serniclaes W (1973) La simultanéité des indices dans la perception du voisement des occlusives. Rap. Act. Inst. Phonétique, Univ. libre Bruxelles, 7/2, 59-67.

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 [sə] 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 patterns 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 s'pose for suppose. In fact, vowel syncope may spill over into more formal styles as in the network news commentary reporting recent "S'preme Court decisions".

In the preceding example, vowel syncope results in a word with two juxtapositioned consonants resembling a dictionary word which does indeed contain a cluster. For example, vowel syncope in support results in the production of s'port which them becomes a possible homonym with sport. Just how listeners identify words containing vowel loss which become homonyms with real words is the 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 confusions 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 tape 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 [sə] followed by [p] or [k], 2) a real word containing an initial cluster consisting of [sp] or [sk], and 3) a word containing an artificially created [sp] or [sk] cluster resulting from vowel syncope. The subjects reported no more difficulty in pronouncing such items as s'port than the other members of the triad, sport and support. Five tokens of each phrase for all subjects were analyzed spectrographicly. 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, these data did not show the systematic reduction in length of /s/ in clusters as opposed to singletons reported by Klatt [5] 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 alternate productions in a trial and error approach. As expected, they also inserted vowels within the real clusters 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 information in the speech stream that was undetected in the studies by Fokes and Bond [4,5]. Of interest also was whether differentiation between real and artificial clusters is an ability restricted to American listeners or whether 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 tape recorded readings of short phrase or sentence triads containing test words 1) with an initial unstressed syllable beginning with [s], 2) a real [sp] or [sk] cluster, and 3) an artificial [sp] or [sk] cluster. Each member of a triad was inserted into the following phrase sets:

On (succumbing, scumming, s'cumbing) at parties.
He (secured, skewered, s'cured) the meat.
The (supplies, splice, s'plies) of tape.
My (support, sport, s'port) 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 unstressed vowel in words containing either the artificial or real clusters. The reduced vowel was present in the test words with the unstressed syllables.

2.2. Subjects

The two groups of subjects were college students: 15 native American English listeners and 10 non-native listeners. The non-native groups' experience with English was limited to academic training in English in their homeland and from two to three 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 headphones 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 56.6% 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 90% and from 86% to 96% for two-syllable words. They identified artificial clusters as two-syllable words from 15% for the non-native speaker to 47% for one of the native speakers. Interestingly, the non-native subjects perceived the proficient non-native speaker's artificial clusters as the target word more often than the native subjects.

Identifications were also lexically dependent; s'cumb was rarely heard as succumb (8%), while s'port and s'cured were identified as two-syllable words 64% of the time. In fact, with the word scum removed from the analysis, identification of the artificial cluster rose to 59% for Speaker Four's productions and to 69% for Speaker 2. Identification 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 measures analysis of variance consisting of one between factor (two listener groups), and two within factors (4 English speakers and word 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 = 2.27,52.11; p<.01); speaker by word triad (F = 36.11; df = 3.26, 75.02, p<.0001); and speaker by listener group by word triad (F = 5.81; df = 3.26, 75.02; p<.0009). There was no listener group by word triad interaction. In determining the source of the interactions, Speaker One was clearly different in that her 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 = 23.35; df = 1, 23; p <.0001); speaker (F = 45.97; df = 2.27, 52.11: p <.0001); and word triads (F = 464.22; df = 1.44, 33.11, p <.0001).

4. CONCLUSIONS

The American native subjects were better able to identify artificial clusters as the target word containing the unstressed initial syllable than the non-natives. This ability cannot be credited to semantic cues only since 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, any one speaker may have used a set of cues which, in turn, may have signaled the intended target word.

In addition, listeners may have the facility of adapting to the peculiarities of individual speakers and their intentions. Apparently listeners are able to perform in this manner even when given a minimal amount of speech data.

5. REFERENCES

[1] Crystal, T. H. and House, A. S., (1988a), Segmental Durations in connected- speech signals: current results, Journal of the Acoustical Society of America, 83, 1553-1573.

[2] Crystal, T. H. and House, A. S., (1988b), Segmental durations in connected- speech signals: syllabic stress, Journal of the Acoustical Society of America, 83, 1574-1585.

[3] Delattre, P. (1966). A comparison of syllable length conditioning among languages, International Review of Applied Linguistics, 4, 184-196.

[4] Fokes, Joann and Bond, Z.S., (1989). S'pose a vowel is lost, Spring Meeting of the Acoustical Society of America, Syracuse, New York.

[5] Fokes, Joann and Bond, Z.S., (1989). Vowel syncope in native and non-native English, American Speech-Language Hearing Association Convention, St. Louis Missouri.

[6] Klatt, D.H. 1975, Voice onset time, frication, and aspiration in word-initial consonant clusters, Journal of Speech and Hearing Research, 18, 686-705.

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

	Native Mean	Native 95% C.I.	Non-native Mean	Non-native 95% C.I.
TRUE CLUSTERS				
S1	99.2	97.9-100	90.4	82.6-98.3
S2	98.9	97.1-100	83.8	74.0-93.0
S3	97.8	96.3- 99.1	79.2	69.0-89.3
S4	99.2	97.9-100	86.7	78.4-95.0
TWO SYLLABLE WORDS				
S1	99.2	98.2-100	86.7	76.9-96.5
S2	99.7	99.1-100	96.7	92.1-100
S3	99.4	98.6-100	95.0	92.3-97.7
S4	100	100 -100	95.4	89.6-100
ARTIFICIAL CLUSTERS				
S1	7.2	3.2-11.3	15.8	7.5-24.1
S2	55.6	48.8-62.3	35.7	29.8-43.5
S3	48.1	39.0-57.1	42.5	28.3-56.8
S4	45.3	37.1-53.5	46.7	32.8-60.6

CENTRAL MECHANISMS OF VOWEL PERCEPTION, CATEGORIZATION AND IMITATION

Inna A. Vartanian, Tatiana V. Chernigovskaya

I.M.Sechenov Institute
Academy of Sciences of the USSR, Leningrad

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 skills are usually associated with the functions of the left hemisphere (LH). The last decades undoubtedly proved the fact of the right hemisphere (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 show that predominant LH or RH influence on information processing is determined by the task factor - either 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: detection, imitation and categorization of speech and complex non-speech samples.
The authors are grateful for the help of prof. N. Svetozarova, Leningrad State University, U.S.S.R. and Dr. K. Ogorodnikova, Bryn Mawr Coll. PA, U.S.A. for the construction and recording of stimuli set. Parts of this paper, under a different title , were presented at the Annual Meeting of the International Neurophyziological Society, San Antonio, Texas, February 1991.

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 sets 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 set. The resulting tape consisted of 24 trials with 3-sec.interval which permitted subjects to record their responses manually or vocally. The stimuli were presented monaurally to the right or the left ear in turn. Reaction time and type of answer were registered automatically. All possible combinations of hands and ears were used. Subjects were asked to give simple vocal or 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 36 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 rythm 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 in space (approaching or moving away). The stimuli were presented monaurally to the left and right ears in quasirandom order. Subjects were instructed to respond monaurally (left or right in different sessions). Reaction time was automatically registered.

3.RESULTS
Subjects turned out to be grouped in two extremes the remaining arranged in between as to their psychophysiological organization. The comparison of the group differences reveals (i) the "reciprocal" character of one of them, i.e. sharply different latent times depending on the stimulation sides, the parameters of the stimuli being identical and (ii) the "synaergic" 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 reactions show 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 movements of the stimulus without phonation. The greatest reaction time was registered when the stimuli were presented to the left ear, while the response was given by the left hand; the least - when the stimuli were 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 stimuli 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 accent-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 acoustic parameters - rythm of pulses, duration, carrier frequency, amplitude shifting, the side of stimulation etc. - in the average-latent time was 1.5 sec. However, it should be emphasized that the usage of "speech-like" criterion increases by 30 per cent when the signal is being addressed to the right hemisphere, i.e. to the left ear. 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 subjects, the dependence of lateralization on the experimental conditions - side of stimulation, type of task, type of answer (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 information 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 levels of perception. LH provides for phonemic encoding and structural analysis of complex acoustic stimuli both in perception and imitation using short-term memory; RH realizes global template recognition. It should be emphasized that perception is language 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 of importance in interpreting mean or normalized data.

5. REFERENCES

CHERNIGOVSKAYA T. (1990), "Modes of consciousness: cultural, functional and neurophysioloical dimensions", *Proc. of Conf. "The Phenominal Mind"*, Bielefeld, ZiF.

CHERNIGOVSKAYA T., BALONOV L., DEGLIN V. (1983), "Bilingualism and brain functional asymmetry", *Brain and Language*, v.20.

CHERNIGOVSKAYA T., DEGLIN V. (1986), "Brain functional asymmetry and neural organization of linguistic competence", *Brain and Language*, v.29.

CHERNIGOVSKAYA T., VARTANIAN I. (1989), "Cerebral asymmetry in speech processing", *Proceed. of "Speech Research '89", Int. Conf. 1989*, Budapest.

DEGLIN V., TRACHENCO O., CHERNIGOVSKAYA T. (1987), "Sound shape of language and cerebral asymmetry", *Proc. of the XI Int. Congress of Phonetic Sciences*, Tallinn.

VARTANIAN I.A. (1987), "The role of speech activity in realization of acoustic perception", *Sensory Systems*, v.1, N.2, (in Russian).

VARTANIAN I.A. ET AL. (1981), "Auditory recognition of complex sounds (psychophysical and clinical aspects)", *Physiol. of Human*, v.7, (in Russian).

VARTANIAN I.A., CHERNIGOVSKAYA T.V. (1980), "Effects of different parameters of acoustic stimulation on estimation of changes of distances from the sound source by humans", *Physiol. Journ. U.S.S.R.*, v.66, (in Russian).

VARTANIAN I., CHERNIGOVSKAYA T. (1991), "Acoustic space and speech perception", *J. of Clin. and Exper. Neuropsychology*, v.13, N 1.

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 excised from conversations with 3 people and their intelligibility was measured. Sentence stress, 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 a 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, or given. 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 contrastive 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 in 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-nonstrident distinction is phonemically contrastive, then, strident sounds might be expected to retain a high level of intelligibility in most contexts.

A phonetic difference that is mainly dependent on phonological space is the leniting of velar stops in English. The only velar consonants in English are oral and nasal stops; so, since /ŋ/ can only be syllable-final, and the acoustic correlates of nasalization are fairly distinctive and distributed over time, leniting /ɡ/ and /k/ 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 ɡ θ s ʃ/.

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 uniqeness 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 ʃ/ are strident; the others are not. 2. /ɡ/ 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 /θ/ resembles /s ʃ/ 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 consonantal fragments will not differ in 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 ʃ/) sounds will be more intelligible than other sounds overall, and later instances will be as intelligible as the first instance.
4. Because we expect /ɡ/ to vary more than /d/, /ɡ/ will be more 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 discussed 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 production of the word; 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 resulting 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 5dB 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 4s, 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 frication 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 2s, 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 comparison. We summarise some 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 goes 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 make-up (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 confounding effect here of prosodic context, since 13 of 15 *new* items are in nuclear position, but 4 of 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. Strident fricatives achieved by far the best scores (/ʃ/: 91%, /s/: 87%), with /d/ and /g/ intermediate (56% and 55%) and /ə/ 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%).

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

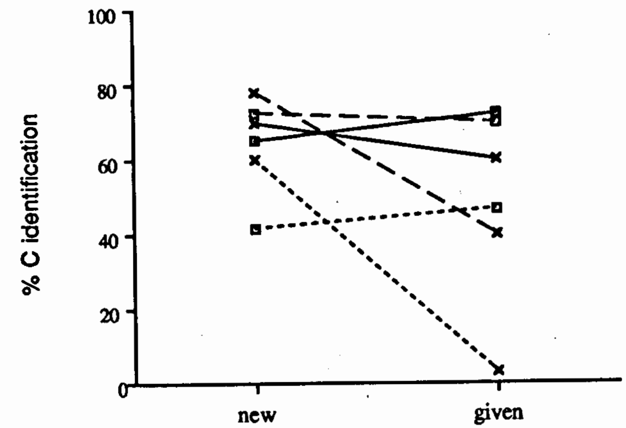


Figure: Consonant intelligibility for /d/-/g/ subset. Different line styles denote the 3 speakers. Crosses show /g/ and squares /d/ identification scores.

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 /g/ but not /d/ (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 intelligibilty 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].

7. REFERENCES

[1] BARD, E.G., LOWE, A.J. & ALTMANN, G.T.M. (1989), "The effect of repetition on words in recorded dictations", *Eurospeech: Proc. 2nd Eur. Conf. on Speech Comm.*, *2*, 573-576.

[2] FOWLER, C.A. & HOUSUM, J. (1987), "Talkers' signalling of 'new' and 'old' words in speech and listeners' perception and use of the distinction", *J. Mem & Lang*, *26*, 489-504.

[3] HAWKINS, S. (1989), "Reconciling trading relations and acoustic invariance", *Eurospeech*, *2*, 682-685.

[4] HUNNICUTT, S. (1987), "Acoustic correlates of redundancy and intelligibility", *Stockholm: KTH. STL-QPSR*, *2-3*.

[5] KOOPMANS-VAN BEINUM, F.J. & VAN BERGEM, D.R. (1989) "The role of 'given' and 'new' in the production and perception of vowel contrasts in read text and in spontaneous speech", *Eurospeech*, *2*, 113-116.

[6] LIEBERMAN, P. (1963), "Some effects of semantic and grammatical context on the production and perception of speech", *Lang. & Speech*, *6*, 172-187.

[7] WARREN, P. & MARSLEN-WILSON, W.D. (1987), "Continuous uptake of acoustic cues in spoken word recognition", *Perception & Psychophysics*, *41*, 262-275.

CENTRAL MECHANISMS OF INTONATION PROCESSING - COMPREHENSION AND IMITATION

Tatiana V.Chernigovskaya, Inna A. Vartanian

I.M.Sechenov Institute
Academy of Sciences of the USSR, Leningrad

ABSTRACT

It is now well recognized that the right hemisphere is concerned with processing of prosodic features of speech - intonation, rhythm and stress. There are however contradictory data concerning linguistic prosody as most of the research involve affective stimuli only. The 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. cuirses) were selectively spared in some groups of aphasics. In 1947 J.Monrad-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 past 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 and revealing communicative centers of the sentence - arbitrary syntactic complexity with meaning differentiating prosody; (iiii)emphatic prosody types, 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 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 latent periods when the stimuli were presented to the left ear than to the right one. Communicatively or syntactically different phrases appeared to be a complex perceptual domain - some 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 - functional 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 specific 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) linguistically balanced stimuli, (iii) new type of procedure - noise for masking the other side of perception, reaction-time measuring, 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.

BALONOV L., DEGLIN V. (1976) Speech and hearing of the dominant and subdominant hemispheres", Leningrad, *Nauka*, (in Russian).
BEHRENS S. (1988), "The Role of the production of linguistic stress", *Brain & Language*, v. 33.
BERNDT R., SALASSO A., MITCHUM CH., BLUMSTEIN S. (1988), "The role of imitation cues in aphasic patient's performance of the grammaticality judgment test", *Brain and Language*, v. 34.
BLUMSTEIN S., COOPER W.E. (1974), "Hemispheric processing of intonation contours", *Cortex*, v. 10, N 2.
BOROD C., KOFF E. (1985), " Channels of emotional expression in patients with unilateral brain damage", *Arch. Neurology*, v. 42.
CHERNIGOVSKAYA T. (1990), "Modes of consciousness: cultural, functiocallosal connections in speech prosody", *Brain and Language*, v. 35.

CHERNIGOVSKAYA T., DEGLIN V. (1986), "Brain functional organization of linguistic competence", *Brain and Language*, v. 29.
CHERNIGOVSKAYA T., VARTANIAN I. (1989), "Cerebral asymmetry in speech processing", *Proceed. of "Speech Research's 89" Intern. Conference*, Budapest.
DEGLIN V., TRACHENKO O., CHERNIGOVSKAYA T. (1987), "Sound shape of language and cerebral asymmetry", *Proceed. of the XI ICPhs*, Tallinn, v. 1.
EFRON R. (1990), "The decline and fall of hemispheric specialization", *Hillsdale*, NJ, Erlbaum.
EMMORY K.D. (1987), "The neurological substrates for prosodic aspects of speech", *Brain and Language*, v. 30.
HEILMAN K., BOWERS D., SPEEDIE L, COSLETT H. (1984), "Comprehension of affective and nonaffective prosody", *Neurology*, v. 34, N 7.
JACKSON H. (1874), " Notes on the physiology of language", *Selected papers*, London.
LEY R., BRYDEN M.P. (1982), " A dissociation of right and left hemisphere effects for recognizing emotional tone and verbal content", *Brain and Cognition*, v. 1, N. 1.
MONRAD-KROHN J.H. (1947), "Disprosody or altered "melody of language", *Brain*, v. 70.
ROSS E.P. (1981), "The aprosodias", *Archives of neurology*, v. 38.
SHAPIRO B.E., DANLY M. (1985), The role of the right hemisphere in the control of speech prosody in propositional and affective contexts", *Brain and Language*, v. 25, N 1.
SVETOZAROVA N. (1987), "Linguistic factors in sentence-stress", *Proceed. of the XI ICPhs*, Tallinn, v.6
WATERSON N. (1987), "Prosodic phonology. The theory and its application to language acquisition and speech processing", *Grevatt & Grevatt*.
ZEIDEL E., CLARKE J.M., SUYENOBU B. (1990), "Hemispheric independence: A paradigm case for cognitive neuroscience", In *A.B. Scheibel and A.J. Wechsler (Eds.) Neurobiology of higher cognitive function* N.Y. Guilford Press.

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

Benoît Jacques*, Maria Amalia Garcia Jurado et Miguelina Guirao****

***Université du Québec à Montréal, Canada, et **Laboratorio de Investigaciones Sensoriales, CONICET, Buenos Aires, Argentine**

ABSTRACT

This paper compares acoustical and temporal cues of /l/ and /r/ in Montreal French and Buenos Aires Spanish with their identification in syllabic context.
Two Argentinian and two French Canadian speakers recorded short sentences in which /l/ and /r/ figure in CV, VC and /a/CV contexts with the vowels /i/, /a/ and /o/. Segments of the waveform were selected for perceptual analysis. It was found that, most of the time, in both languages, the liquids are percieved as modulations of the intensity or the timbre of the contiguous vowel and that they cannot be identified unless the selected segment contains three or more cycles of that vowel. The modulations 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 présentent des observations sur les propriétés acoustiques et perceptuelles de ces consonnes en espagnol (Guirao et Rosso [4], Garcia Jurado, Guirao et Rosso [3]) et en français (Chafcouloff [1,2], Santerre [5,6,7], Tousignant [8]).
Nous nous proposons de comparer les principales caractéristiques acoustiques et temporelles de /l/ et /r/ en relation avec leur identification en contexte syllabique, en particulier de 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 /r/ en contextes de CV, VC et /a/CV avec les voyelles /a/, /i/ et /o/. L'onde complexe obtenue par ordinateur 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 train d'ondes 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 amputés 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 syllabe. 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 syllabe, 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": une plus grande durée de la voyelle modifiée avec une durée plus brève de la voyelle libre équivaut à la combinaison opposée d'une durée plus brève de la voyelle modifiée avec une durée plus longue de la voyelle libre.

l l/a a

Figure 1: Esp. [la]

a a/r r

Figure 2: Esp. [ar]

3.1. Syllabes avec /l/

3.1.1. Positions initiale vs finale

En espagnol, lorsque /l/ est en position initiale, ou observe un segment de basse amplitude dont le timbre est celui d'une voyelle neutre suivi d'un autre de plus grande amplitude qui correspond au noyau vocalique.
La transition entre les deux segments est abrupte avec /a/, moins abrupte avec /o/ et graduelle avec /i/. En français, la syllabe avec /a/ présente les mêmes caractéristiques que sa correspondante espagnole. Par contre, avec /o/, on a observé chez un locuteur une diphtongaison, /l/ étant perçu comme un [i] et la syllabe, [io], au lieu de [lo]. Enfin, lorsque /i/ 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 [i] prolongé.
En position finale, /l/ espagnol montre une transition graduelle entre les voyelles /a/ et /o/ et la liquide et celle-ci se perçoit comme une modulation d'amplitude de celles-là. En contexte de /i/, 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 [i] suivi d'une pulsion sans timbre défini, elle-même suivie d'un autre [i] d'amplitude plus faible que le premier. En français, on observe une transition graduelle des trois voyelles avec la consonne, à cela près que, dans les contextes de /a/ et de /i/, s'ajoute une vélarisation, /il/ et /al/ étant perçus respectivement [iɑ] et [ao]. A noter que le /o/ précédant /l/ dans les segments français est la voyelle brève ouverte comme dans *bol*, non le /o/ fermé long de *pôle*.
A l'exception des contextes 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 msec en position initiale et près du double en position finale.

3.1.2 Position intervocalique /a/CV

En espagnol et en français, les séquences /ala/ et /alo/ reproduisent les mêmes phases que celles déjà observées dans les combinaisons /al/, /la/ et /lo/ et les segments critiques sont de durée comparable en espagnol, légèrement supérieure en français. Quant à la séquence /ali/ de l'espagnol, si on peut bien reconnaître le passage de la première voyelle à la liquide, il n'en va pas ainsi pour le passage de celle-ci à /i/, parce que, dès le début de son déroulement, /l/ prend le timbre de cette voyelle. Pour ce qui est de la même suite en français, on constate que /a/ se transforme en [e] par harmonisation partielle avec la voyelle /i/ et que la liquide prend elle aussi ce nouveau timbre: /ali/ devient [eei].

3.1.3. Les durées minimales

TABLEAU 1
Durées minimales des segments permettant de reconnaître la liquide /l/ (msec)

	Informateurs			
	argentins		canadiens	
	1er	2e	1er	2e
/la/	65	68	50	43
/li/	117	178	*140	115
/lo/	70	60	75	89
/al/	97	175	100	116
/il/	*187	105	161	234
/ol/	125	78	98	132
/ala/	123	124	104	121
/ali/	110	145	129	110
/alo/	140	140	169	175

*Superposition totale de la voyelle et de la liquide.

Les portions de durée du tableau 1 occupées par la voyelle libre variant entre 30% et 60% pour la position initiale. Des exceptions s'observent dans le contexte de /i/: cette voyelle occcupe en effet 73% de la durée du segment espagnol de 178 msec et il y a un cas de recouvrement complet des deux sons en français. Pour la position finale, la proportion de voyelle libre se situe entre 25% et 48%. La durée minimale nécessaire pour identifier la liquide est plus grande lorsque celle-ci est en position finale. En outre, pour chacune de ces deux positions, à une exception près, c'est le contexte de /i/ qui montre les durées les plus longues. Toutefois, en position intervocalique, il y a équilibre dans les durées vocaliques avant et après la liquide parce que chacune des voyelles fournit un appui favorisant la reconnaissance de celle-ci.

3.2. Syllabes avec /r/

En général, /r/ espagnol se réalise comme une interruption ou un silence dans le segment vocalique. Ceci est surtout vrai en position initiale de syllabe. En position finale, en effet, cette consonne peut parfois se réaliser comme une modulation d'intensité ou de qualité de la voyelle précédente. En français, /r/ peut présenter des vibrations gutturales faibles, observables dans les tracés acoustiques, mais pas toujours perçues à l'audition. Il peut aussi présenter des formants sans vibrations. Dans les deux cas, le résultat au plan perceptuel est une modulation de la voyelle adjacente. Il peut également se réaliser comme une fricative, à l'instar du /r/ parisien. Enfin, en finale, il peut diphtonguer la voyelle qui précède.

3.2.1. Positions initiale vs finale

La réalisation de /r/ initial espagnol suivi de /a/ et /o/ se caractérise par la production d'une voyelle d'appui dont la durée ne peut être inférieure à trois cycles. Cet élément vocalique est suivi d'une interruption d'une durée non inférieure à 15 msec. Avec /i/, l'appui vocalique se réduit à une simple pulsion suivie de l'interruption, de sorte que /ri/ est perçu [pri] ou [bri]. En français, devant /a/ et /o/, /r/ initial est perçu comme une voyelle de basse intensité et de timbre indéfini. Toutefois, dans un cas avec /a/, il apparaît comme un son guttural similaire à une fricative sonore. Les deux contextes de /i/ ne permettent pas de percevoir la consonne: dans l'un, il y a réduction de surface et toute la syllabe est disparue; dans l'autre, c'est un [i] sans modulation qui est perçu tout au long de la syllabe et l'identification phonologique de /ri/ ne se fait qu'au niveau du mot.

En espagnol, /ar/ et /or/ se réalisent de la même façon que lorsque /l/ termine la syllabe, à savoir un segment vocalique dont l'intensité baisse graduellement, se transformant en un autre segment de durée égale ou supérieure. Ce changement se produit dans un intervalle de 40 à 60 msec (segment critique). Par contre, on trouvera après /i/ une interruption d'environ 25 msec suivie d'une voyelle brève de même durée, ou une interruption suivie d'une pulsion. En français, comme en espagnol, /ar/ et /or/ se présentent comme des segments vocaliques modulés. Dans un contexte de /a/ et les deux contextes de /o/, il y a diphtongaison de la voyelle et /r/ se réalise comme un [o] ou un [u]. Enfin, précédé de /i/, le /r/ montréalais se perçoit comme une voyelle centrale ou comme un [e].

3.2.2. Position intervocalique /a/CV

En espagnol, on voit se reproduire pour les suites /ara/ et /aro/ les mêmes phases que dans les suites déjà observées de /ar/, /ra/ et /ro/. En ce qui concerne /ari/, la présence des deux voyelles adjacentes est nécessaire pour l'identification de la consonne, celle-ci étant réduite à un bref intervalle silencieux. En français, /r/ se transforme en élément vocalique de basse intensité. Dans /ara/, cet élément prend le timbre des deux voyelles; dans /ari/, il prend le timbre de la seconde, tandis que dans /aro/, chez un informateur, il adopte le timbre de la première voyelle et, chez l'autre, il se perçoit comme un [u].

3.2.3. Les durées minimales

TABLEAU 2
Durées minimales des segments permettant de reconnaître la liquide /r/ (msec)

	Informateurs argentins 1er	2e	Informateurs canadiens 1er	2e
/ra/	90	73	52	90
/ri/	62	78	*80	–
/ro/	95	114	137	90
/ar/	97	99	145	280
/ir/	98	110	155	90
/or/	80	144	197	130
/ara/	136	68	141	150
/ari/	100	97	181	106
/aro/	100	120	283	157

*Impossibilité de séparer /r/ de la voyelle.

Dans les deux langues, lorsque /r/ est en position initiale, la portion du segment occupée par la voyelle, excluant les appuis vocaliques des /r/ espagnols, varie entre 33% et 66%, avec une moyenne autour de 50%. En position finale, cette portion varie entre 50% et 80% pour l'espagnol (moyenne de 56%) et entre 13% et 40% seulement pour le français. Il faut toutefois noter que ces pourcentages faibles relevés en français sont des fractions de segments relativement longs, puisque /r/ final se présente essentiellement comme une modulation graduelle de la voyelle qui précède. Pour cette raison, les durées minimales nécessaires pour l'identification de /r/ final sont plus longues que celles requises pour reconnaître /r/ initial, à plus forte raison lorsque celui-ci est fricatif, comme c'est le cas pour le segment /ra/ de 52 msec. En espagnol cette différence reliée aux positions tend à se limiter au contexte de la voyelle /i/.

En ce qui concerne la position intervocalique, il faut noter la durée très longue du segment /aro/ chez le premier informateur canadien: la liquide ayant pris le timbre de la voyelle précédente, elle ne peut être perçue et reconnue avant le début de l'articulation de /o/.

4. REMARQUES GENERALES

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

Une liquide ne peut jamais être identifiée au plan perceptuel sans la présence d'au moins une partie de la voyelle adjacente.

/l/ est semblable dans les deux langues et se présente comme une modulation d'amplitude et parfois de timbre de la voyelle adjacente.

/r/ est différent, puisqu'en espagnol, il se présente surtout comme une interruption précédée d'un appui vocalique, alors qu'en français il prend des aspects divers, incluant ceux d'une fricative, bien que le plus fréquent que nous ayons observé soit une modulation de la voyelle adjacente analogue à celle produite par /l/.

La durée minimale nécessaire pour percevoir les liquides est plus longue, lorsque celles-ci sont en position finale de syllabe que lorsqu'elles sont à l'initiale. Ceci s'observe dans les deux langues et dans les trois contextes vocaliques en ce qui concerne /l/; dans le cas de /r/, cette différence s'observe aussi dans les trois contextes en français, mais en espagnol, elle tend à se limiter au contexte de /i/.

Le contexte de /i/ est différent des contextes de /a/ et de /o/, lesquels sont similaires entre eux. La voyelle /i/ et la liquide tendent à se superposer davantage, ce qui augmente la durée 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, /l/ et /r/ peuvent présenter des similitudes en position finale, à cause de l'absence d'interruption dans le /r/. Les deux liquides demeurent toutefois bien différentes en position initiale. En français, les similitudes ou les différences que peuvent présenter entre elles les deux liquides sont indépendantes des positions.

5. REFERENCES

[1] CHAFCOULOFF. M. (1972), *Comparative phonetic study of /l/ in American English, French, German and Spanish*, Université de Provence: Institut de Phonétique.

[2] CHAFCOULOFF. M. (1980), "Les caractéristiques acoustiques de [j, ɥ, w, l, r] en français", *Travaux de l'Institut de Phonétique d'Aix, 7*, 7-56.

[3] GARCIA JURADO, M. A., GUIRAO, M., ROSSO, E. A. (1989), "La influencia de la duracion en la identificacion de las consonantes liquidas", *III Congreso Internacional de El Español de America, Valladolid, España, Julio 3-9 1989*

[4] GUIRAO, M., ROSSO, E. A. (1987), "Acoustic-phonetic analysis of Spanish /l/", *114th Meeting of the Acoustical Society of America, Miami, Florida, USA, nov 16-20 1987, Journal of the Acoustical Society of America, suppl 1, 82*, CCC4.

[5] SANTERRE, L. (1979), "Le /r/ montréalais en régression rapide", *Protée XII, II*, 117-131.

[6] SANTERRE, L. (1982), "Des [r] montréalais imprévisibles et inouïs", *Revue québécoise de linguistique, 12, 1*, 77-96.

[7] SANTERRE, L. (1989), "Peut-on juger de la production par la perception? (ou faut-il en croire ses oreilles?)" *Mélanges de phonétique générale et expérimentale offerts à Péla Simon, 2*, Publications de l'Institut de Phonétique de Strasbourg, 735-755.

[8] TOUSIGNANT, C. (1987), *La variation sociolinguistique, modèle québécois et méthode d'analyse*, Montréal, Presses de l'Université du Québec.

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

Vincent J. van Heuven & Marc Ch. Dupuis

Dept. Linguistics/Phonetics Laboratory,
Leyden University, The Netherlands

ABSTRACT

This paper examines the perceptual effects of first and second order anticipatory coarticulation in V1CV2-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 influences from a non-adjacent 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. Öhman [2] showed that part of the behaviour of the formant transition movements in V1 toward C in V1CV2 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. [1]). 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 well above chance level on the basis of the transitions from, or into, the neighbouring 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 V1CV2 sequences under optimal conditions.

Vowels located in the central area of the traditional two-dimensional vowel diagram should be more prone to adjustment under 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 /i,a,u/ (hypothesis 1).

We predict further that effects of coarticulation 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 vice versa. We have used stimuli prepared from fragments 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 particularly strong 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 /i,a,u/. The targets, such as tafel 'table' or koepel 'dome', were monomorphemic 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 realised 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 plosive at the beginning of the target word. The resulting 72 sentences were copied on a test tape in nine series of eight sentences. In each series the order of the stimulus sentences was randomized. The interstimulus interval was fixed at 7 s (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 is affected by the type of preceding vowel (V1) and the accent pattern over V1/V2 is shown in table I.

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

	RESPONSES FOR	
	C	V2
V1 accented		
V1= /i/	65	32
/u/	62	38
/a/	85	38
schwa	80	41
V1 unaccented		
V1= /i/	80	38
/u/	64	32
/a/	87	44
schwa	82	50
Overall	76	39

C-identification

The overall correct identification score for C was 76%, which is way above chance (=33%). Obviously, the type of V1 played an important role in the identification of C. The deleted consonants were, on the whole, identified best from preceding /a/. The overall effect of V1 on consonant identification was strongly significant [X^2 (3) = 185.5, $p < .001$]. While subjects identified C significantly better from schwa (81% correct) than from /i/ (73% correct) or /u/ (63% correct), the difference in scores between /a/ (86% correct) and schwa contexts was likewise found to be significant [X^2 (1) = 10.2, $p < .01$]. Our first prediction, viz. that stops are better identified in the environment of preceding schwa than after point vowels, was therefore not quite confirmed by the overall results of VC-coarticulation.

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% [X^2 (1) = 15.5, $p < .001$].

V2-identification

The vowels /a/ and especially /u/ were identified well above chance while identification of /i/ was not. The total 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 is substantial [X^2 (3) = 32.8, $p < .001$]. Identification is significantly better when V1 is schwa (45% correct) than when V1 is /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, is essentially correct for vowel-onto-vowel coarticulation.

Examining effects of stress distribution on the identification of V2 we observe that scores were generally higher for stimuli in which V1 was unaccented and V2 was accented (41% correct) than for stimuli in which the distribution of stress was reversed (37% correct) [X^2 (1) = 5.9, $p < .05$]. We conclude that hypothesis 2, whereby 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% <u>versus</u> 50% correct) between the two accentuation conditions can be observed in contexts where V1 was schwa [X^2 (1) = 8.3, $p < .01$]. The value of 50% correct identification for V2, measured in unaccented schwa contexts, exceeds all other values. This result shows that, as far as identification of V2 is 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 correctly identified segments from 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 found that percent correct scores were of equal magnitude in the environment of preceding /schwa, a/, which were both significantly better than the environments /i/ and /u/.

As concerns the role of V1 with respect to the identification of V2, our results clearly 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 both 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, predicting additivity of stress distribution and vowel type, stands.

Our experiment is the first to show convincingly that perceptual effects of anticipatory coarticulation from-vowel-onto-vowel are not necessarily restricted to immediately adjacent segments. When conditions are carefully chosen, the perceptual effect of the second order vowel-onto-vowel effect can be substantial. Clearly, the reason why other researchers have by and large failed to uncover convincing perceptual effects of vowel-onto-vowel coarticulation ([1,3,4,5,6]), lies in their infelicitous choice of stimulus material. Notice, in this context, that our optimal condition (predicting an accented V2 from a preceding unstressed 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 used outside the laboratory in 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 # 300-161-023

5. REFERENCES

[1] BENGUEREL, A.P. & S. ADELMAN (1975). Coarticulation of lip rounding and its perception, in A. Cohen & S.G. Nooteboom (eds.), <u>Structure and process in speech perception</u>, Berlin: Springer Verlag, 283-293.

[2] DUPUIS, M.CH. (1988). Perceptual effects of phonetic and phonological accommodation, Doct. diss. Leyden University.

[3] KUEHN, D.P. & K.L. MOLL (1972). Perceptual effects of forward coarticulation, <u>J. Speech Hearing Res. 15</u>, 654-664.

[4] LEHISTE, I & L. SHOCKEY (1972). On the perception of coarticulation effects in English VCV syllables, <u>Working Papers in Linguistics 12</u>, Ohio State University, 78-86.

[5] MARTIN, J.G. & H.T. BUNNELL (1981). Perception of anticipatory coarticulation effects, <u>J. Acoust. Soc. Am. 69</u>, 559-567.

[6] MARTIN, J.G. & H.T. BUNNELL (1982). Perception of anticipatory coarticulation in vowel-stop consonant-vowel sequences, <u>J. Exp. Psychol.: Human Percept. Perform. 8</u>, 473-488.

[7] ÖHMAN, S.E.G. (1966). Coarticulation in VCV utterances: spectrographic measurements, <u>J. Acoust. Soc. Am. 39</u>, 151-168.

EFFECT OF VOWEL QUALITY ON PITCH PERCEPTION

I.Raimo, O.Aaltonen, 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 members 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 [a, æ], 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 cataract type synthesis. All vowels had the same input amplitude configurations and the following formant structures (Hz) and method-dependent relative amplitudes:

	F1	F2	F3	dB
a	700	1100	2500	+5-6
æ	650	1700	2500	+3-4
u	300	600	2500	+1-2
i	300	2250	2850	+0-1

The durations of all vowels were 23 cs. Five F0 levels (1-5) were used. For Level 1, F0 was 104 Hz initially, reaching 114 Hz after 9 cs 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, 160 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 51% correct judgments. Below, the percentages of selections as the higher 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 the choir, 1320 judgments.)

	Students	Choir
æ - u	79- 9 (12)	64-24 (13)
a - u	75- 8 (17)	56-27 (17)
i - u	68- 9 (23)	56-27 (18)
æ - i	49-24 (27)	45-37 (18)
æ - 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 - i	23-18 (59)	17-34 (49)
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 [u] 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 *1-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			Choir		
	ks1	*ks2*	*AL*	*ks1*	*ks2*	*AL*
a	4.5	5.4	-40	3.1	5.2	-10
æ	4.0	5.7	-16	4.0	5.7	-13
i	3.2	5.2	-4	3.6	4.9	-13
u	0.8	3.6	+7	4.1	5.4	-2
Mean	3.1	5.0	-14	3.7	5.3	-9

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 **[a]**. 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 **[æ]** 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 *s* values explain both all vowel-specific pitch differences and the TOEs in our data. Besides, pitch discriminability (indicated by *ks*) 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 æ a]**. 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 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

[1] GANDOUR, J. and WEINBERG, B. (1980), "On the relationship between vowel height and fundamental frequency: evidence from esophageal speech", *Phonetica, 37,* 344-354.

[2] HELLSTRÖM, Å. (1985), "The time-order error and its relatives: mirrors of cognitive processes in comparing", *Psychological Bulletin, 97,* 35-61.

[3] HELLSTRÖM, Å. (1989), "What happens when we compare two successive stimuli?", in G. Ljunggren & S. Dornic (Eds.), *Psychophysics in Action,* Berlin: Springer.

[4] HONDA, K. & FUJIMURA, O. (forthcoming), "Phonological vs. biological explanations - intrinsic vowel pitch and phrasal declination", *Vocal Fold Physiology,* Proceedings of the 6th Vocal Fold Physiology Conference, Stockholm 1989.

[5] LEHISTE, I. (1970), *"Suprasegmentals"*, Cambridge: The MIT Press.

[6] LIBERMAN, A. & MATTINGLY, I. (1985), "The motor theory of speech perception revised", *Cognition, 21,* 1-36.

[7] SAPIR, S. (1989), "The intrinsic pitch of vowels: theoretical, physiological, and clinical considerations", *Journal of Voice, 3* , 44-51.

[8] VILKMAN, E., AALTONEN, O., LAINE, U. & RAIMO, I. (forthcoming), "Intrinsic pitch of vowels - a complicated problem with an obvious solution?", *Vocal Fold Physiology,* Proceedings of the 6th Vocal Fold Physiology Conference, Stockholm 1989.

[9] VILKMAN, E., RAIMO, I. & AALTONEN, O. (1991), "Is subglottal pressure a contributing factor to the intrinsic F0 phenomenon?", in this publication.

THE PERCEPTION OF SILENT-CENTER SYLLABLES IN NOISE

Susan L. Hura

University of Texas, Austin, Texas

ABSTRACT

The perception of vowel-less /b/-vowel-/t/ 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 wavefrom 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 simple target 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-/t/ syllables in the carrier phrase "I say the word /bVt/ somemore," 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 for full and SC versions of a syllable at the same S/N. Six S/N were created for each syllable type for each /bVt/ 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 /bVt/ word they heard on a preprinted answer sheet. There were a total of 252 items in the test—it lasted 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 $F(2,162)=4.22$ at $p<.025$; S/N is also significant for $F(5,162)=8.01$ 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.

	-6	-3	0	3	6	9
full	60.3	74.2	85.0	87.7	90.4	89.6
SC1	53.1	48.8	76.5	86.5	75.4	80.0
SC2	56.5	58.1	60.4	73.5	78.9	84.2

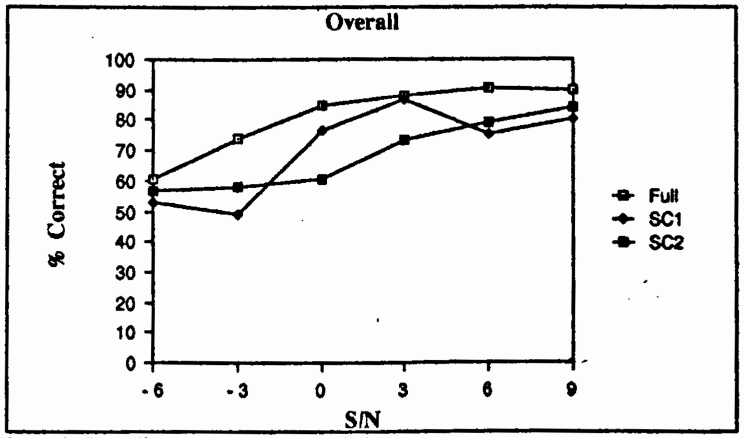


Figure 1: Overall percent correct by syllable type.

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 redundacies 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

[1] ANDRUSKI, J. & NEAREY,T. (forthcoming), "Comparing the perception isolated vowel and /bVb/ syllables in hybrid silent center stimuli" *JASA*.

[2] DIEHL,R., McCUSKER,S. & CHAPMAN, L. (1981), "Perceiving vowels in isolation and in consonantal context" *JASA, 69*, 239-248.

[3] JOOS, M.(1948). "Acoustic Phonetics" *Language Supplement, 24*, 1-136.

[4] LINDBLOM, B. & STUDDERT-KENNEDY, M. (1967), "On the role of formant transitions in vowel recognition" *JASA, 42*, 830-843.

[5] NEAREY, T. and ASSMANN, P. (1986), "Modelling the role of inherent spectral change in vowel identification" *JASA, 80*, 1297-1308.

[6] PARKER, E. & DIEHL, R. (1984), "Identifying vowels in CVC syllables: effects of inserting silence and noise" *Perception & Psychophysics, 34*, 369-380.

[7] STRANGE, W. (1989), "Dynamic specification of coarticulated vowels spoken in sentence context" *JASA,85 (5)*, 2135-2153.

[8] STRANGE, W., VERBRUGGE, R., SHANKWEILER, D.& EDMAN, T. (1976), "Consonantal context specifies vowel identity" *JASA, 60*, 213-224.

MINIMAL DURATION FOR PERCEPTION OF FULL-SPECTRUM VOWELS

Rudolf Weiss
Western Washington University
Bellingham, WA, USA

W. S. Brown, Jr.
University of Florida
Gainesville, FL, USA

Shaw N. Gynan
Western Washington University
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 waveshapes (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 bursts of vowels at set-time intervals with manipulation of F_1 and F_2 frequencies [2,3]. Other studies have involved masking techniques [5] and still others have dealt with vowel formant transitions for vowel vs. consonant vs. semi-vowel identification [4]. Most vowel perception experiments share in common the fact that they use synthesized vowels with manipulations of F_1, F_2 and/or F_3 relative to each other in frequency, band-width 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 waveshapes (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 waveshapes (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 Liebermann [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 (F_0 = 100 Hz ± 2 Hz) and a female speaker (F_0 = 201 Hz ± 3 Hz). These vowels were digitized using the MacSpeech Lab II/MacAudio II hardware/software program. A sampling rate of 44 KHz 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 seg-mented 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 female utterances are given below:

Male:	F_0	F_1	F_2
[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:	F_0	F_1	F_2
[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 F_0 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. Analog 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 SB-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.

<u>Male</u>		<u>Female</u>	
[o]	2.1%	[o]	7.6%
[u]	5.6%	[a]	8.0%
[i]	5.8%	[i]	9.7%
[a]	5.9%	[e]	24.0%
[e]	14.9%	[u]	35.5%

It is obvious from the above statistics that the most difficult male vowel to categorize was [e], 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. Errors are indicated as a percentage.

TABLE 1: PERCEPTION ERRORS OF COMPARABLE M/F VALUES

<u>ms</u>	[i]		[e]		[a]		[o]		[u]	
	M	F	M	F	M	F	M	F	M	F
5	13.1	21.9	18.4	37.7	8.7	9.6	0	10.5	7.0	51.7
10	3.5	1.7	11.4	14.9	8.7	7.8	4.3	7.0	7.8	59.6
20	4.3	0.8	9.6	12.2	3.5	11.4	0	3.5	4.3	35.0
30	6.1	7.0	23.6	11.4	5.2	7.0	4.3	4.3	7.0	28.0
40	1.7	1.7	11.4	11.4	3.5	1.7	1.7	5.2	1.7	18.4

The study shows that in general errors in perception increase as the vowel duration decreases. An exception is the male [o] which posed no difficulty for the listeners even at the shortest duration of 1/2 wave cycle (5 ms). Recognition levels for the shortest durations were as follows:

Male (tokens for 1/2, 1 and 2 cycles)

5 ms:	90.4% (81.6-100%)
10 ms:	92.9% (88.6-96.5%)
20 ms:	95.7% (90.4-100%)

Female (tokens for 1/2, 1-5 cycles)

2 ms:	74.6% (49.2-84.3%)
5 ms:	74.7% (48.3-90.4%)
10 ms:	79.7% (40.4-93.0%)
15 ms:	84.8% (57.9-98.2%)
20 ms:	87.4% (65.0-99.2%)
25 ms:	92.8% (84.3-99.2%)

For context independent recognition of vowels the male voice obviously yields the best response. With the exception of [e] all vowels could be truncated to one wave form (10 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 cue. 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 IASCP, University of Florida.

7. REFERENCES

[1] BLADON, A., (1983), "Two-formant models of vowel perception: shortcomings and enhancements", *Speech communication* 2, 305-313.

[2] CHISTOVICH, I., et al. (1987). "Interval of spectral information accumulation in perception of non-stationary vowels", *Proceedings XIth ICPhS*, 1, 262-265.

[3] CHISTOVICH, L.A. (1985), "Central auditory processing of peripheral vowel spectra", *JASA*, 77 (3), 789-805.

[4] DANILOFF, R.G. (1985), *"Speech science"*, San Diego: College Hill Press, 146 pp.

[5] LIEBERMANN, P. and S. BLUMENSTEIN (1988), *"Speech physiology, speech perception, and acoustic phonetics"*, New York: Cambridge University Press, 175 pp.

[6] REPP, B.H., et al. (1979), "Categories and context in the perception of isolated steady-state vowels", *Journal of experimental psychology, human perception and performance*, 5 (1), 129-145.

[7] WEISS, R. (1976), "The role of perception in teaching german vowels to american students", *Proceedings of the IVth international congress of applied linguistics*, 3, Stuttgart: Hochschulverlag, 513-523.

[8] WEISS, R. and H.H. WAENGLER (1975), "Experimental approach to the study of vowel perception in German," *Phonetica*, 32 (3), 180-199.

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 confirmed observations that French /y/ is usually pronounced as an /u/-like vowel by English speakers, and as an /i/-like vowel by Portuguese speakers. The results of a perceptual test in which the same speakers were asked to identify a set of synthetic stimuli constituting a high vowel continuum (from /i/ to /u/) revealed that the accented pronunciations of French /y/ by English and Portuguese speakers are accounted for 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 3 high vowels /i/, /y/, and /u/, native speakers of languages whose inventory contains only the 2 high vowels /i/ and /u/ find it difficult at first to pronounce the target vowel /y/. 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 /y/, 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 /y/ by speakers whose native languages contain only the 2 high vowels /i/ and /u/ 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 /y/), 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 25 to 32) took part in the experiment: speakers of Canadian English, who have been observed to replace French /y/ 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 /a/ in different consonantal contexts. In the perceptual task, English and Portuguese subjects were asked to identify as /i/ or /u/ 3 sets of randomized synthetic stimuli constituting a high vowel continuum (from /i/ to /u/), with one set consisting of isolated vowels and the other 2 of vowels in the environments /b/__ and /d/__ 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 Klatt'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 in 100 Hz steps, with F1 held constant at 250 Hz. F3 was held constant at 2212 Hz for stimuli with F2 values between 500 and 1800 Hz, and was calculated according to the following formula for stimuli with F2 values above 1800 Hz: F3 = 1.4 x (F2-220) [6]. FØ decreased linearly from 120 Hz at the start of the vowel to 100 Hz at its end. The 21 members of each continuum were presented 10 times each in random order for forced-choice identification. They were low-pass filtered at 4800 Hz and delivered binaurally through TD-149 earphones.

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 7-point scale: 1= /i/ or /i/-like vowel; 2 = vowel between /i/ and /y/, but closer to /i/; 3 = vowel between /i/ and /y/, but closer to /y/; 4 = /y/ or /y/-like vowel; 5 = vowel between /y/ and /u/, but closer to /y/; 6 = vowel between /y/ and /u/, but closer to /u/; 7 = /u/ or /u/-like vowel. On the basis of this scale, a score between 1 and 4 indicates a vowel between /i/ and /y/, and a score between 4 and 7 a vowel between /y/ and /u/. The stimuli were presented on-line on a Zenith 286 microcomputer, by means of software developed at the University of Alberta, and delivered binauraly through TD-149 earphones.

When they were not successful in repeating French /y/ as /y/ or an /y/-like vowel, Portuguese speakers repeated it 95% of the time as /i/ or an /i/-like vowel (generally a lax variant thereof), or as a vowel described by the 3 French judges as falling between /i/ and /y/. They repeated French /y/ as /u/, an /u/-like vowel, or even a vowel between /y/ and /u/ 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 /y/ as /y/ or an /y/-like vowel, they were found to repeat it as /u/ or an /u/-like vowel (a lax variant thereof), or as a vowel between /y/ and /u/ 92% of the time, and as an /i/-like vowel or a vowel between /y/ and /i/ 8 % of the time. Their mean score for these non-/y/ productions was 5.01. These results support observations that Portugese speakers generally replace French /y/ 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 /u/ 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 ranging between 1500 and 2100 Hz, which are identified as /y/ by French speakers, are most of the time labeled as /u/ by English speakers and as /i/ by Portuguese speakers.

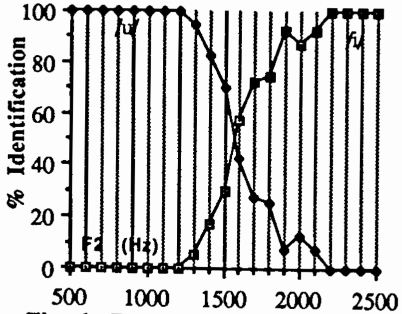


Fig. 1: Portuguese Identification Functions

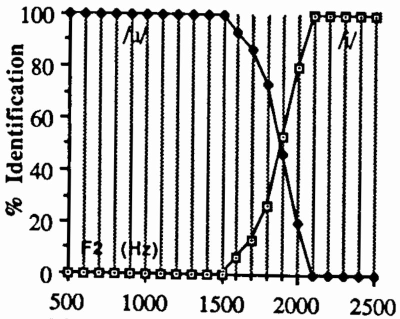


Fig. 2: English Identification Functions

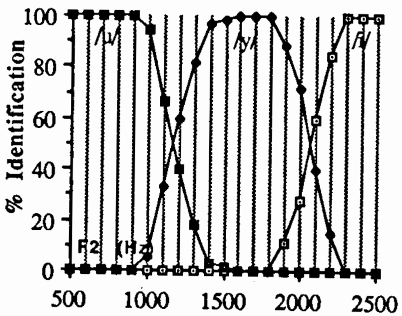


Fig. 3: French Identification Functions

Although this comparison of the labeling functions by the three groups of subjects is of interest inasmuch as it reveals how these three groups of subjects divide the high vowel continuum in their respective languages, it should not form the basis for an explanation of the phenomenon of interlingual identification. When called upon to imitate French /y/, L2 learners do not have access to French categorization functions, but only 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 /y/ obtained from production data to the L2 learners' identifications functions. The average value of F2 for French /y/ has been given as 1850/1900 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 presented to the English and Portuguese subjects in the imitation task of the experiment reported here was 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 /i/ category for Portuguese speakers, and within the /u/ category for English speakers.

4. DISCUSSION

4.1. The parallelism between the results of the imitation task and those of the perceptual task appear to support the hypothesis that accented pronunciations of L2 sounds by untrained speakers may be perceptually motivated. It suggests that, in early stages of L2 learning, learners perceive L2 sounds in terms of their L1 phonological systems, through the process of "equivalence classification" [3]. Thus, they may classify separate L2 phonemes as acoustically different realizations of the same L1 category, even if they perceive the acoustic differences in question. Once assigned to that category, the intended target speech sound is actualized according to their L1 phonetic realization rules. In the case of Portuguese speakers, most tokens of French /y/ fall within the bounds of the Portuguese /i/ category. Once assigned to this perceptual category, such tokens are imitated by Portuguese speakers in such a way that they are perceived by French speakers as belonging to their own /i/ or /y/ categories (see Fig. 3). On the other hand, for English speakers, most tokens of French /y/ fall within the bounds of the English /u/ category. Once assigned to this category, such tokens are imitated by English speakers in such a way that they are perceived by French speakers as belonging to their own /u/ or /y/ category. The fact that intended French /u/, as pronounced by English speakers, is often perceived by French speakers as /y/ is further evidence that English speakers assign both French /y/ and /u/ to their single /u/ category; English /u/ being characterized by higher F2 values than its French counterpart [4], its realizations cover a range which straddles the French /y/ and /u/ categories.

4. 2. The results of this study further provide evidence that there exist differences in the way different languages divide the high vowel continuum. The results of the perceptual test for English speakers agree with the findings of Stevens et al. [8] who, in their crosslanguage study of vowel perception, observed a peak in the discrimination functions for both English and Swedish speakers, as one passed from the unrounded /i/ to the rounded /y/. Because the English speakers were able to perceive the acoustic differences between /i/ and /y/ in spite of the fact that there is no distinction between an /i/ and an /y/ category in English, the authors concluded that some natural perceptual boundary must exist between these two vowels. The identification functions represented in Fig. 2 indicate that, although English has only two high vowel categories labelled /i/ and /u/, the perceptual boundary between these two categories nearly coincides with the perceptual category between /i/ and /y/ 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 occured not because of a natural perceptual boundary in the region, but because the stimuli being discriminated belonged to 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 two high vowels /i/ 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 a reflection of some basic property of the auditory mechanism.

5. REFERENCES

[1] DELATTRE, P. (1948), "Un triangle acoustique des voyelles orales du français", *The French Review*, 21:6, 477-484.

[2] DELATTRE, P. (1969), "An acoustic and articulatory study of vowel reduction in four languages", *International Review of Applierd Linguistics*, 7:4, 295-325.

[3] FLEGE, J. E. (1988), "The production and perception of foreign language speech sounds", in H. Winitz (Ed.), *Human Communication and its disorders: A review-1988*, Norwood, New Jersey: Ablex Publishing.

[4] FLEGE, J. E., and J. HILLENBRAND (1984), "Limits on phonetic accuracy in foreign language speech production", *Journal of the Acoustical Society of America*, 76:3, 708-721.

[5] KLATT, D. (1980), "Sofware for a cascade/parallel synthesizer", Journal of the Acoustical Society of America, 67, 971-995.

[6] NEAREY, T. M. (1989), "Static, dynamic and relational properties in vowel perception." *Journal of the Acoustical Society of America*, 85:5, 2088-2113.

[7] O'SHAUGHNESSY, D. (1982), "A study of French spectral patterns for synthesis", *Journal of Phonetics*, 10, 377-399.

[8] STEVENS, K.N., A.M. LIBERMANN, M. STUDDERT-KENNEDY, and S.E. G. OHMAN (1969), "Crosslanguage study of vowel perception", *Language and Speech* 12:1, 1-23.

[9] WALZ, J. (1979), *The early acquisition of second language phonology*, Hamburg, Germany: Helmut Buske.

This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada (410-89-0854).

UNDERSTANDING DISFLUENT SPEECH: IS THERE AN EDITING SIGNAL?

R. J. Lickley, R. C. Shillcock and E. G. Bard.

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 claimed 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- you'd really need about eight ...'

Example 2: False Start:

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

It is all to 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) every self-correction site [...] is marked by a phonetically identifiable signal placed at the right edge of the expunction site ...'*
> ([1] p128)

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 expunction site'* (ie 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 imitated versions 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 3:

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 to ...'

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 (*it's*), 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 gated 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 creaky 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

[1] HINDLE, D. (1983), "Deterministic Parsing of Syntactic Non-Fluencies", *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*

[2] LABOV, W. (1966), "On the Grammaticality of Everyday Speech", *Paper presented at the Annual Meeting of the Linguistic Society of America.*

INDIVIDUAL VARIABILITY IN THE PERCEPTION OF CUES TO AN INITIAL BA-PA VOICING CONTRAST

V. Hazan and B. Shi

Dept of Phonetics & Linguistics, University College London, U.K.

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. Continua 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 ranges. 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 /ei/ environment than in an /a/ 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 -20ms to +70 ms in nine steps *and* change in the F1 onset frequency.

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

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

2.1. Natural edited stimulus continua

Recording were made of tokens of /pa/ 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 100Hz to facilitate editing in 10ms steps.

The creation of natural-edited stimulus continua was done on a mini-computer using a "cut and paste" technique. For the Ba/VOT continuum, the vowel portion from the [ba] token was appended to the burst and aspiration portion from the [pʰa] token. For the creation of stimuli with VOT between 70 ms and 5 ms, the aspiration was progressively deleted from the vowel end

in 10 ms steps. For stimuli with negative VOTs, the prevoicing portion was cut out from the [ba] stimulus and appended to the initial burst. The same process was carried out for the Pa/VOT continuum, except that, in this case, both the burst/aspiration and vowel portions were taken from the [pʰa] 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 resynthesis through a 4 kHz bandwidth, software parallel formant synthesiser was performed. Further modifications to the syntheses were then made on the basis of comparisons between the natural and synthetic spectra on a Kay digital spectrograph 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 10 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 [1]).

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. The mean boundaries obtained were 13.5 ms (s.e. 5.1) for the natural edited condition, 13.3 ms (s.e. 6.2) for the stylised synthesis 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 synthesis condition was compared with labelling of the other full-cue conditions. 83 % of subjects showed significant deviances at the 0.001 level (deviances greater than 26.1) between the copy-synthesised and stylised synthesis continua. Significant deviances were only found for 44% of subjects when the natural edited and stylised synthesis stimuli were compared and the range of deviances 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 deviances 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 deviance (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 deviance between the two conditions with condition deviances 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 deviance measures obtained. However, more homogeneous results were obtained with natural-edited stimuli than with copy-synthesised stimuli.

7. REFERENCES

[1] HAZAN, V. and ROSEN, S. (1991) Individual variability in the perception of cues to place contrasts in initial stops. *Perception and Psychophysics*, vol.49, 2.
[2] LISKER, L and ABRAMSON, A. (1970). The voicing dimension: some experiments in comparative phonetics. *Proc. of the 6th ICPhS*, Prague, 1967 (Academia, Prague), 563-567.

[3] HOLMES, W. (1989) Copy synthesis of female speech using the JSRU parallel formant synthesiser. *Proc. of Eurospeech '89* (Paris), 513-516.
[4] SHI, B. and HAZAN, V. (in press) Effect of stimulus type on the labelling of a /ba/-/pa/ voicing contrast. *Speech, Hearing and Language, UCL Work in progress*, vol.5.

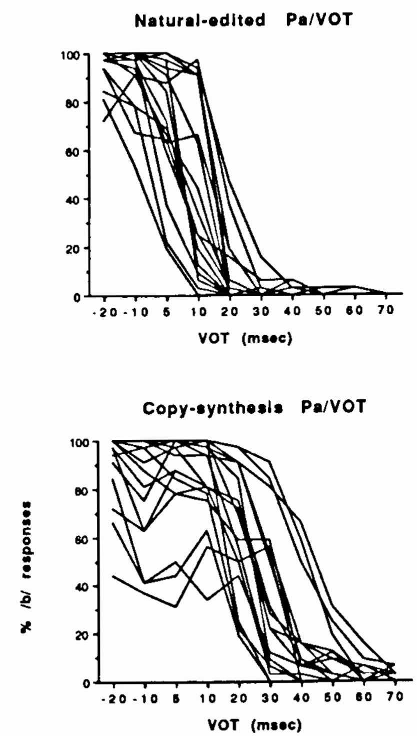


Figure 2: Individual labelling functions for the two Pa/VOT conditions.

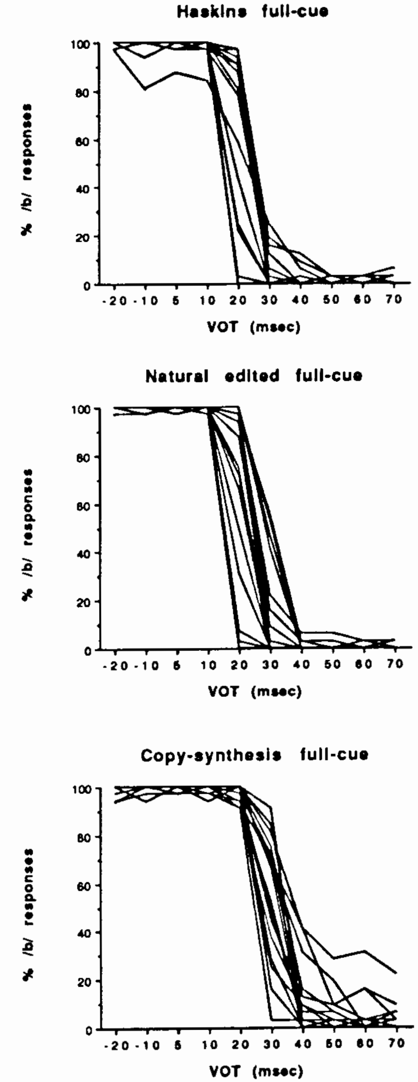


Figure 1: Individual labelling functions for the three full-cue conditions.

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=18) 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[3]. 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 20 vowels. 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

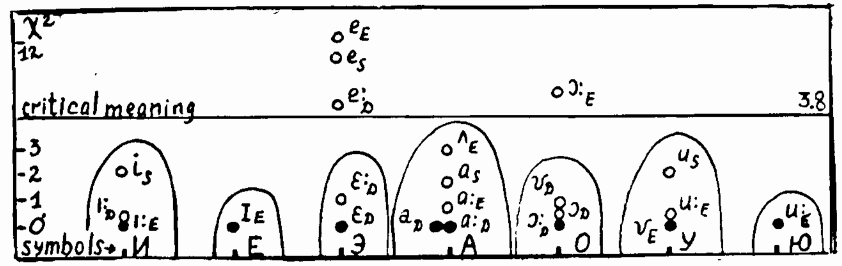


Fig.1. The distribution of the listeners' answers (by the χ^2 criteria) on the vowels which give the most reliable identification.

within CV and VC syllables. The number of listeners was 43, and the total number of vowels and syllables was 161. In all the experiments we received from the listeners 1947 answers. The listeners knew neither of the above mentioned languages, and they also didn't know the sounds of what languages they were listening to. They were asked to write what they heard by means of the letters of the Russian alphabet.

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 are more numerous than that of Russian (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 20 types 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 ⟨Э⟩ more frequently than other letter-symbols to mark the vowels they were presented to. The ⟨Э⟩ space includes 6 vowels: /e_E, e_S, e:_D, ɛ:_D, ɛ_D, æ_D/. The spaces of the Russian ⟨А⟩ and ⟨О⟩ include 5 vowels: ⟨А⟩ - /a:_E, ʌ_E, a_S, ɑ_D, a:_D/; ⟨О⟩ - /ɔ_E, ɔ:_E, ɔ_D, ɔ:_D, ʊ_D/. Then come the Russian ⟨И⟩ and ⟨У⟩ including 3 vowels each: ⟨И⟩ - /i:_E, i_S, i:_D/; ⟨У⟩ - /ʊ_E, u_S, u:_D /. The most narrow are the spaces of ⟨Е⟩ and ⟨Ю⟩ including only one vowel each: ⟨Е⟩ - /I_E/; ⟨Ю⟩ - /u:_E/. One can also see from the Table, on which vowels the greatest number of answers marked by one letter-symbol falls. These are the following vowels: /i:_E/, /I_E/, /ɛ_D/, /ɑ_D/, /a:_D/, /ɔ:_D/, /ʊ_E/, /u:_E/. n = 8. All of them are reliably identified by the Russian listeners. Fig. 1 shows the results of comparison of answers' distributions on those vowels which were marked in most cases by one letter-symbol. The difference between the distributions of answers for each pair of vowels was estimated by means of χ^2 criteria. The comparison of the distributions shows that the listeners are able to distinguish not 8 but at least 12 vowels. As it is seen, 4 vowels are placed higher than the critical meaning of the χ^2 criteria is. 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, secondary, the remoteness of non-native vowels from the centre formed by the native vowel in the linguistic consciousness of the Russian speakers.

Let's analyse another group of vowels. While identifying these vowels the listeners do not take unanimous decisions. The vowels are: /æ$_E$/,/ɔ$_E$/, /ɜ:$_E$/,/O$_S$/,/I$_D$/,/y:$_D$/, /y$_D$/,/œ$_D$/,/ø:$_D$/ (see Table 1). The task which the listeners had to fulfil 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 with the unknown vowel stimili. Let's consider the vowels which differ only in one step of openness articulatory similar 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 /ɜ:/ are identified mostly as ⟨EЭ⟩(70% and 60% of all the answers correspondingly). Comparison of the answers' distributions shows that these two vowels form one perceptual sphere for the speakers of Russian and they may be placed in the space of ⟨EЭ⟩ in Fig.1. German /I/ is identified in the 50% of all the answers as ⟨И⟩ and in the 50% as ⟨EЭ⟩. Thus, it may be placed in the space of ⟨ИEЭ⟩ on Fig.1. English /ɔ/ and Spanish /O/ are identified as the Russian ⟨O⟩ and ⟨A⟩ (50% of all the answers correspondingly). 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 sphere /ɔO/ which can be placed in the space of ⟨AO⟩ on Fig.1. While identifying the German /y, y:, œ, ø:/ the Russian listeners use from 8 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 character 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 allophones in the position between or after palatalysed consonants [2].

The analysis of the distribution of inadmissible answers shows that the vowels /y/ and /œ/ are most similar in the space of phonetic features from the viewpoint of the Russian speakers. The vowels /y:/ and /ø:/ are well distinguished by the listeners.

Table 1. Identification of English, Spanish and German vowels by the native speakers of Russian.

Vowels / symbols	English											Spanish					German														
	i:	I	e	æ	a:	ʌ	ɔ	ɔ:	ʊ	u:	ɜ:	i	e	a	o	u	i:	I	e:	ε	ε:	a	a:	ɔ	ɔ:	ʊ	u:	y	y:	œ	ø:
Э		10	50	30							30	3,8	51,6			1,8		35	66	91	82			4,8		3,6		18,4		51,3	20
А			10	5	90	80	35							85,7	74	2,2						100	100								
О							55	60	15		5		2	2	49,8	5								81	86	75		2,9		14	6
И	85	5										67,5	11	0,4			83	45	8,6										8,6		
У	10			10	5			40	70	20					6,7	58								4,8	14	14	69	17	20	9	2,4
Ю										75																	19		37	2,9	4,1
Е	5	60	25	40							30	1,7	7,8					14	11	9	12							5,7		11,4	
Ы		10							5	5	5	11,3	8,2		2,1	7	7									24			5,7		
Ё											25																	3,6	2,9	8,6	31
ОА					5	10	10																								
УЭ																										3,6	12		2,9		
ЕЭ																					4									2,9	2,4
ИЭ																		6	6									2,9	5,7		2,4
ИЮ																													5,7		
ЕО																														5,7	2,4
ЮУ																												2,9			
УЦ																													2,9		
ОЕ																															6,5
ИО																															4,8
ИУ																															4,8

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 18;[12 (Fig.1) /I$_D$/,/æɜ:/,/ɔO/,/yœ/,/y:/, /ø:/. 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 18 and they form vertical spheres in Fig.1; 2)The vowels which are placed in a perceptual space, formed in a linguistic consciousness of the Russian speakers by several symbols: /i:$_D$/-⟨ИEЭ⟩,/æɜ:/-⟨EЭ⟩,/ɔO/-⟨AO⟩; 3)the vowels which are not placed in a perceptual space ("alien"to it). These are the German front rounded vowels.

6. REFERENCES

1.BONDARKO,L., VERBITSKAYA,L., ZINDER,L., PAVLOVA,L.(1966) "The distinguished sound units of the Russian speech". In:"Mechanisms of speech production and perception of complex sounds",M.,165-179 (in Russian).

2.BELYAKOVA,G.,LEBEDEVA,G., OGORODNIKOVA,K.(1989)"About the perception of front rounded vowels".In:Experimental phonetic analysis of speech,L.,36-46(in Russian)

3.TCHERNOVA,E.,BELYAKOVA,G. MALINNIKOVA,T.(1986)"The investigation of perception of cardinal vowels by the speakers of Russian" Zeitschrift für Phonetik und Kommunikationsforschung,39,N4

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[1]

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-blending) to determine whether such hypothesized units as the onset, rime or coda 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.[2] 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-blending task that could be group administered. Since the main question of interest related to the direction of the primary bonding 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 coda of the other, as illustrated by the following example: SAN1 + CIM1 —> (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 —> (a) TI5 (b) PA5. 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-blending 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 transliteration 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 illiterate Portuguese speakers. (See [9,10] for further discussion of problems with the notion of the phoneme 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 the 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 the Korean letters for SA placed on top and the letter for N placed below it, thus implying an organization like (SA)N rather than S(AN). In addition, Youn has recently conducted an informal word-blend production task, whose results to date support this analysis (see [11]). A Korean version of the forced-choice word-blending task is now under way to firm up these preliminary findings, but the results of that 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 inversions at all. When a similar problem 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 ... indicates the location of the pause):
(a) /mɛ...lən/ (where /l/ is treated as the onset of the second syllable),
(b) /mɛl...ən/ (where /l/ is the coda of the first syllable), or (c) /mɛl...lən/ (where /l/ is ambisyllabic). In an extensive pilot study, this task was presented to 95 undergraduate 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 or re-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 every literate speaker presumably knows. The purpose of the present investigation, therefore, was to establish whether any general preference could be found that was independent 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, depending on the morphological structure of the words involved. The phonemic string MILI in standard Korean, for example, is ambiguously syllabified in the orthography as MI/LI (when it means 'in advance') or as MIL/I (when it means 'wheat + nom'), where a slash is used here to show the location of the break between the syllable-sized *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 great majority of whom grew up in the general Seoul area. The results were as follows:[3] (1) The clearest cases involved single consonants that are restricted phonotactically to syllable-initial position, such as /č/ (as with SA-/CANG [1.00]), or to syllable-final position, such as /ŋ/ (as in PANG-/I [1.00]). (2) The results were also very clear for consonant clusters, where the preferred break occurs between them. This result was virtually unanimous if this break corresponded with the spelling (as in CHENG-/SO [.99] and KUK-/SU [.98]), but remained the majority choice even when the orthography put the break after the second consonant (e.g., AN-C/A [.74] and KAP-S/I [.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.[4] (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.)

MI-/LI	(.91/.95)	vs. MI-L/I	(.91/.27)
A-/NI	(.83/1.00)	vs. A-N/I	(.81/.20)
I-/PYENG	(.66/.97)	vs. I-P/YENG	(.45/.25)
CE-/KE	(.55/.95)	vs. CE-K/E	(.55/.63)
SO-/KA	(.53/.97)	vs. SO-K/A	(.52/.25)

If the post-vocalic break position was unambiguously supported by the spelling for such consonants, the effect was, of course, maintained and even enhanced (e.g., I-/MOKI [.89/.94]), but if an unambiguous spelling break was located after the consonant, a major shift again occurred in that direction (as in KI-L/I [.48/.54]). (Notice that supplying the meaning had little effect for these two words, which was the general trend for the non-ambiguous items throughout.) The single outlier pair among the ambiguous strings was KO-/KI (.90/1.00) and KO-K/I (.87/.25), which in the 'no meaning' condition (first numbers) both yielded the kind of results expected for <u>non</u>-ambiguous strings, as discussed in (1)-(3) above. (Compare also the second set of figures in the first column above, where disambiguation was achieved by supplying the meanings.) Given the very high frequency and familiarity of the word KO-/KI (meaning 'meat'), we suspect that our subjects were simply insensitive to the spelling ambiguity here under the 'no meaning' condition (KO-K/I is the nominalized form of a relatively rare word meaning 'musical piece').

4. CONCLUSIONS

Our attempt to expand the experimental exploration of syllable structure to languages beyond English has been slowed by the fact that new experimental techniques have had to be developed in nearly all cases. Nevertheless, the following preliminary results can now be reported: (1) Korean syllables appear to be of the left-branching or head+coda type, challenging the universality of the onset+rime strategy; (2) the syllables of the Chinese dialects (in this case Taiwanese) continue to resist experimental attempts to subanalysis, casting further doubt on the universality of the phoneme as a basic unit (cf. [10]); and (3) Korean speakers show a decided preference to divide V-/C/V and VC-/C/V sequences at the positions marked by hyphens, even though their orthography permits syllable breaks at all four of the positions marked by slashes.

NOTES

[1]The research reported here was supported by a research grant from the Social Sciences and Humanites 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], this volume), except that the linkages in Korean, as expected, are to the following vowel, rather than to the preceding one.

REFERENCES

[1] TREIMAN, R. (1983), "The structure of spoken syllables: evidence from novel word games," *Cognition 15*, 49-74.

[2] TREIMAN, R. (1986), "The division between onsets and rimes in English syllables," *Journal of Memory and Language 25*, 476-491.

[3] DOW, M.L. (1987), *"On the psychological reality of sub-syllabic units,"* Ph.D. dissertation, University of Alberta, Edmonton.

[4] DOW, M.L. & B.L. DERWING (1989), "Experimental evidence for syllable-internal structure." In R. Corrigan, F. Eckman & M. Noonan (Eds.), *"Linguistic categorization,"* Amsterdam: John Benjamins, 81-92.

[5] DERWING, B.L. & T.M. NEAREY (1991), "The 'vowel-stickiness' phenomenon: three experimental sources of evidence," in this volume.

[6] TREIMAN, R. & C. DANIS (1988), "Syllabification of intervocalic consonants," *Journal of Memory, and Cognition 27*, 87-104.

[7] READ, C., Y-F. ZHANG, H-Y. NIE, & B-Q. DING (1986), "The ability to manipulate speech sounds depends on knowing alphabetic writing," *Cognition 24*, 31-44.

[8] MORAIS, J.L. CARY, J. ALEGRIA & P. BERTELSON (1979), "Does awareness of speech as a sequence of phones arise spontaneously?," *Cognition 7*, 322-331.

[9] DERWING, B.L., T.M. NEAREY & M.L. DOW (1986), "On the phoneme as the unit of the 'second articulation'," *Phonology Yearbook 3*, 45-69.

[10] DERWING, B.L. (in press), "Orthographic aspects of linguistic competence," In M. Noonan, P. Downing & S. Lima (Eds.), *"Linguistic aspects of literacy,"* Amsterdam & Philadelphia: John Benjamins.

[11] YOUN, Y.B. (1990), "Arguments against the universality of the onset/rime division," *Sogang Working Papers in Linguistics 4*, 93-104.

LA PHONÉTISATION DU CASTILLAN

CABRERA C., CONTINI M. et BOË L-J.

Institut de la Communication Parlée, URA CNRS n° 368
Grenoble, France

ABSTRACT

Our project was the establishment of a grammar for the automatic phoneticization of Spanish. By examining a lexicon of 65.000 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 phoneticization 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 hispànique, 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 multilangue 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. Conçu 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

syntaxiques suivants :

- L'unité linguistique sélectionnée est la chaîne graphémique
- Déclaration d'ensembles de natures diférentes 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 réecriture 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'insertion de commentaires dans la grammaire bornés par !

L'intérêt de TOPH réside dans l'accès à des traces 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, outre les transcriptions attachées à chaque entrée, 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 référons 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 déviantes ou supplémentaires par rapport aux règles mentionnées dans les travaux déjà signalés cela 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'édification 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" = (y, w)

b) séparateur de mots
"#" = (-,., .:, ;,",!)

c) "except : i" = (articulad, angular, unívoc, áxic, auricular, atómic, ocinética, odegradable, odegradación, odinámica, offsica, ograf, ográfic, ógrafo, ología, ológico, ólogo, oluminiscencia, omasa, omecánica, ometría, opsia, oquímico, osfera, osíntesis, oterapia, otico, otita, otropismo, óxido, al, ofita,os, ozoo, alin, ato, ante, ogloso, oide,able, abilidad, enio, enal, edro, ásico, ar, ángulo)

Cet ensemble d'exceptions nous permet d'écrire la règle :
("#"+b,br,h,tr,v) +i+ ("except : i") = [i]
sachant que la règle générale (majoritaire)

est :
("consonne") +i+ ("voyelle sauf i") = [j]

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, autrement dit sans tenir compte des exceptions ou des emprunts, la phonétisation du castillan se résume à une soixantaine de règles élémentaires. A titre illustratif, nous nous limiterons 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 (entigrecerse).
Ainsi si l'on considère le trait régulier, un graphème comme "g" sera traité par 3 règles:

+g+ (e,i) = [x]
("#", n) +g+ = [g]
+g+ = [ɣ]

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

("#"+gro,buldo) +g+ ("#") = [g]
("#"+zigza,iceber,basi) +g+ ("#") = [x]
("#"+ban,campin,smokin,bumeran,boomeran,rin,puddin,pin,pon,parkin,marketin,gon,dumpin,dopin) +g+ ("#") = [ ]
("#"+tun) +g+ (steno) = [ ]
("#"+rémin) +g+ (ton) = [ ]
("#"+gan) +g+ (ster+ismo,#) = [ ]
("#"+copyri,bri) +g+ (ht) = [ ]
("#"+neglig) +g+ (é) = [ʃj]
("#"+sufra) +g+ (is+t,m) = [ɣ]
("#"+he) +g+ (elia+n,nismo) = [ɣ]
("#"+per) +g+ (ola) = [g]
("#"+lori) +g+ (a) = [g]
("#"+ideolo) +g+ (o) = [g]
("#"+enti) +g+ (recerse) = [g]
("#"+cat) +g+ (ut) = [g]
+g+ (e, i) = [x]
("#",n) +g+ =[g]
("#") +g+ ("#") = [xe]
+g+ = [ɣ]

Les règles ont été testées sur une base de données conséquente et notamment un dictionnaire de 2500 entrées, implémenté sur HYPERCARD (Macintosh) contenant 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 98%.

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 constatons que ces apports lexicaux désorganisent quelque peu le phonétisme de cette dernière ou du moins n'obéissent pas aux règles de prononciation standard. Cependant quelquefois ils semblent avoir été pratiquement totalement assimilés par la langue (pour les plus anciens) et nous obtenons alors deux prononciations possibles pour une même unité lexicale, une se fondant sur le phonétisme castillan et l'autre conservant les traits de la langue d'origine (bridge, chauvinismo, chauvinista). Materiau vivant, la langue nécessitera pour son étude la constante réactualisation de nos règles ainsi que le renouvellement des lexiques établis.

RÉFÉRENCES

[6] ALCINA FRANCH & MANUEL BLECUA (1988), *Gramática española*. Editorial Ariel, Barcelona, 277-401.
[2] AUBERGÉ & al. (1987), TOPH : un outil de phonétisation multilangue, *Bulletin de l'Institut de Phonétique de Grenoble*, Vol 16, 155-176.
[3] *Gran Diccionario de la lengua española*. (1989), S.G.E.L., Madrid.
[4] NAVARRO TOMÁS (1970), *Pronunciación española.*, Decimoquinta edición, Publicaciones de la revista de filología española, Madrid.
[5] QUILIS & FERNÁNDEZ (1969), *Curso de fonética y fonología españolas*, Cuarta edición, C.S.I.C., Madrid.
[1] SANTOS & al. (1984), Real time text to speech conversion system for spanish, *IEEE-ICASSP*, San Diego, 1593-1596.

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 7854

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 internal 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 pattern. 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? Can 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 terms "stress" is often avoided and is replaced by the partial synonym "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 intonation 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 "minor 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 locations. 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 con-

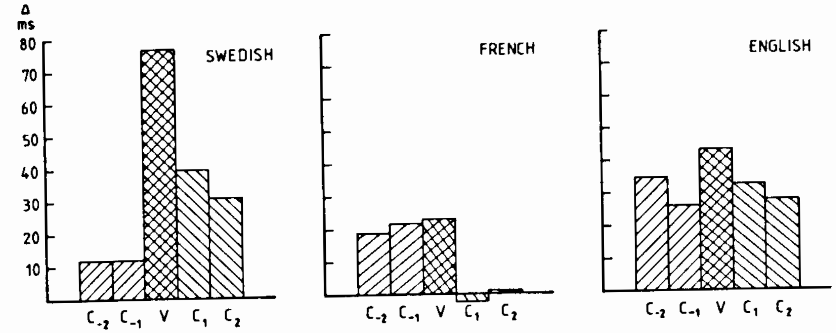


Fig.1. A comparison of stress induced segmental lengthening (measured duration minus unstressed reference) in Swedish, French and English. The cross-hatched columns represent stressed vowels.

trasts 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 Swedish and English. In French the particular distribution is very much dominated by two-phoneme syllables, essentially of 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 to plotting syllable durations against number of phonemes. For Swedish unstressed syllables we find

$$d = 10 + 50m \qquad (1)$$

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

$$D = 57 + 77m \qquad (2)$$

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

$$D = 81 + 52m \qquad (3)$$

Now, if we compare the Swedish and the French data with respect to the same number of phonemes per stressed syllable, 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-

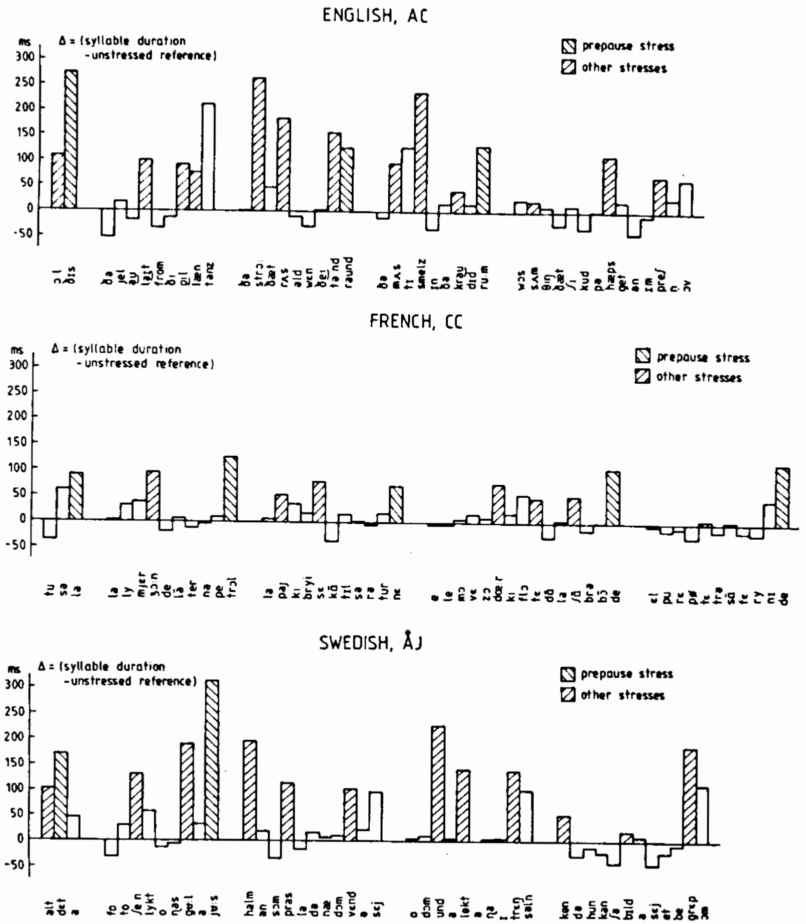


Fig.2. Syllable durations minus unstressed reference values within a sentence in Swedish and in French and English translations.

stressed segmental durations of the phonemes contained. Stressed syllables are cross-hatched.

Again we note the smaller stressed/unstressed contrast in French than in Swedish or English. In all three languages the verb phrase - "she could perhaps get an impression of" - at the end of the sentence is deemphasized as seen by the negative values of the normalized syllable durations, indicating increased tempo. In all three languages there is a definitive prepause lengthening, affecting unstressed as well as stressed syllables. There is no phrase initial shortening of unstressed syllables, at least not for French, but unstressed syllables immediately following a nonterminal stress in French tend to be shortened. Occasionally, a phonologically unstressed syllable is lengthened. This is found in the last syllable of the

word "musty" and the first syllable of the word "flottaient", which could be considered as an instance of "accent d'insistance", also seen in the first syllable of "cela". Also the noun "lumière" is somewhat lengthened, supporting the following adjective "jaune". Some of these traits are of course speaker specific but still of some general interest.

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

[1] Abercrombie, D. (1967), *Elements of General Phonetics*, Edinburgh University Press.

[2] Dauer, R. M. (1983), Stress-timing and syllable-timing reanalyzed, *Journal of Phonetics*, **11**, 51-62.

[3] Delattre, P. (1969), The general phonetic characteristics of languages. Final report OEC-0-9-097701-0775/014, US Dept. of Health, Education and Welfare. University of California, Santa Barbara.

[4] Fant, G. and Kruckenberg, A. (1989), Preliminaries to the study of Swedish prose reading and reading style, *STL-QPSR* **2/1989**, 1-83.

[5] Fant, G., Kruckenberg, A. and Nord, L. (1991), "Durational correlates of stress in Swedish, French and English", Proceedings of the Second Seminar on Speech Production, Leeds, May 1990. To be published in *Journal of Phonetics*.

[6] Pike, K. (1946), *Intonation of American English*, . Univ. of Michigan Press, Ann Arbor.

[7] Touati, P. (1987), *Structures prosodiques du suédois et du français*. Travaux de l'institut de linguistique de Lund, Lund University Press.

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 (a) 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 consonant 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 both 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 with regard 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/, /s/ 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 had 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) containing 4 cases each of initial /p,b,s,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/Fr (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 spectrograms 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 striation for a following vowel; stop 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 striations 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/Fr, Fr/Eng and Eng/Eng. Eng/Eng stops have few negative VOTs, Fr/Eng rather more, and Fr/Fr the most. The results for Sp/Eng speakers differ according to the subject. SP1 produced only prevoiced (i.e. negative VOT) stops in both languages. SP2 produces both short lag VOT and prevoiced 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 prevoiced stops in Spanish.

- In /b/ in carrier sentences, both the English and French subjects' performance is characterised by a good deal of variability. On the whole, the Fr/Fr 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 Fr/Eng or Eng/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 labial approximants. In the Sp/Eng conditions, both subjects consistently produce stop closures, but with variable timing of voicing, producing both fully voiced and partially devoiced tokens of /b/. SP2 produces both fully voiced and partially devoiced 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 /b/ 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 *ns* reflect cases where either speaker error or measurement uncertainty forced exclusion of a token.

Mean VOT in /b/ in isolated words -- separate means given for +ve and -ve VOTs)

Fr/Fr			Fr/Eng		
Fr1	19(8/5)	-71(20/13)	Fr1	21(6-17)	-74(-/1)
Fr2	13(5/5)	-72(14/5)	Fr2	11(2/13)	-96(31/6)
Fr3	24(15/16)	-44(24/4)	Fr3	16(4/18)	-74(-/1)
Fr4	17(-/1)	-80(22/15)	Fr4	11(2/7)	-72(42/8)
Fr5	12(3/8)	-62(22/9)	Fr5	-	-124(12/6)
Sp/Sp			Sp/Eng		
Sp1	-	-119(20/20)	Sp1	-	-119(23/20)
Sp2	14(3/4)	-52(10/15)	Sp2	14(3/17)	-57(11/3)
Sp3	10(-/1)	-48(21/17)	Sp3	26(22/9)	-65(32/11)

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

Fr/Fr	CD	MV	A	B	C	Fr/Eng(CD)	MV	A	B	C
Fr1	74(14/20)	74(14/20)	-	-	20	151(84/20)	137(40/19)	1	1	18
Fr2	91(30/20)	90(18/20)	-	3	17	166(70/20)	140(43/17)	3	6	11
Fr3	67(7/20)	61(14/20)	-	9	11	99(15/20)	83(19/20)	-	7	13
Fr4	103(22/20)	100(24/20)	-	3	17	186(69/20)	98(28/20)	-	16	4
Fr5	93(21/19)	71(29/19)	-	11	8	220(79/16)	110(57/16)	0	10	6
Sp/Sp						Sp/Eng				
Sp1	-	-	-	-	-	113(26/11)	92(61/7)	4	3	4
Sp2	61(14/11)	54(14/11)	-	4	7	100(18/9)	60(21/9)	-	7	2
Sp3	66(19/6)	66(19/6)	-	-	6	102(20/7)	85(26/7)	-	3	4

Mean consonant duration, medial voicing and summary of timing patterns for /z/ in isolated words

Fr/Fr	CD	MV	A	B	C	Fr/Eng(CD)	MV	A	B	C
Fr1	140(51/20)	146(55/14)	6	2	12	168(43/20)	163(38/10)	10	1	9
Fr2	124(26/20)	98(46/16)	4	8	8	140(35/20)	109(64/18)	2	9	9
Fr3	122(18/20)	66(47/19)	1	11	8	124(14/20)	49(40/18)	2	14	4
Fr4	134(29/20)	43(32/20)	-	19	1	133(37/20)	38(43/16)	4	15	1
Fr5	150(37/20)	158(34/17)	3	-	17	142(48/19)	134(60/19)	-	2	17
Sp/Sp						Sp/Eng				
Sp1	-	-	-	-	-	124(50/20)	106(66/13)	7	3	10
Sp2	-	-	-	-	-	60(13/20)	56(17/16)	4	1	15
Sp3	-	-	-	-	-	128(25/20)	87(62/8)	12	4	4

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

Fr/Fr	CD	MV	A	B	C	Fr/Eng(CD)	MV	A	B	C
Fr1	140(42/20)	140(42/20)	-	-	20	165(30/20)	159(40/16)	4	1	15
Fr2	141(35/20)	131(43/20)	-	3	17	160(37/20)	146(49/19)	1	3	16
Fr3	107(12/20)	101(23/20)	-	6	14	121(15/20)	96(40/20)	-	6	14
Fr4	154(40/20)	128(52/20)	-	9	11	135(20/20)	116(40/17)	3	13	4
Fr5	124(27/20)	96(41/20)	-	9	11	188(37/20)	173(55/20)	-	14	6
Sp/Sp						Sp/Eng				
Sp1	-	-	-	-	-	124(18/12)	56(-/1)	10	1	1
Sp2	-	-	-	-	-	84(14/11)	76(21/10)	1	2	8
Sp3	-	-	-	-	-	148(9/12)	48(22/9)	3	9	-

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 being found across all the subjects, 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/Fr 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 consonants 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, it 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 means *and* 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 inherent variability in phonetic targeting. This work is now in progress.

5. REFERENCES

[1] DOCHERTY G. J. (1991) *The Timing of Voicing in British English Obstruents.* Dordrecht:Foris

[2] FLEGE, J. (1991) "Age of learning affects the authenticity of voice-onset time in stop consonants produced in a second language" *J.A.S.A., 89*:395-411

[3] FLEGE, J. & PORT, R. (1981) "Cross-language phonetic interference: Arabic-to-English" *Lg. & Sp., 24*:125-146

[4] KEATING, P. (1988) "The window model of coarticulation" *UCLA WPP, 66*:104-123

[5] LOFQVIST, A., & YOSHIOKA, H. (1984) "Intrasegmental timing: laryngeal-oral coordination in voiceless consonant production" *Speech. Comm., 3*:279-289

[6] PORT, R. & MITLEB, F. (1980) "Phonetic and phonological manifestations of the voicing contrast in Arabic-accented English" *Research in Phonetics, 1*:137-165 (Dept. of Linguistics, Indiana University).

INSTRUMENTAL PHONETIC FIELDWORK TECHNIQUES AND RESULTS

Peter Ladefoged

Linguistics Department, UCLA, Los Angeles, U.S.A.

ABSTRACT

Phoneticians can now take 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 electroglottograhic 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 20-24,000 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 cumbersomely and tediously using tape recorders. But they are trivial, normal operations on any computer equipped with a means for digitizing and editing recorded sounds.

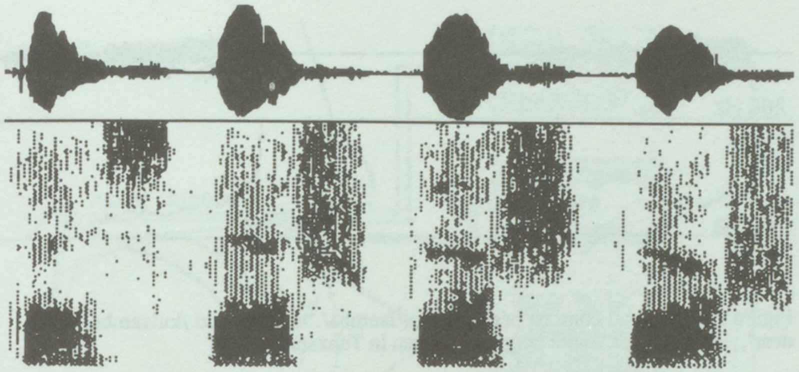


Figure 1. A spectrogram of [kos̪, pos̺, poʃ, poʂ] 'money, milk, language, clan name' in Toda, made under field conditions.

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 laminal sound. In the last word, the lowering of the third formant is probably due to the sublingual cavity that is formed by raising the tip 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.

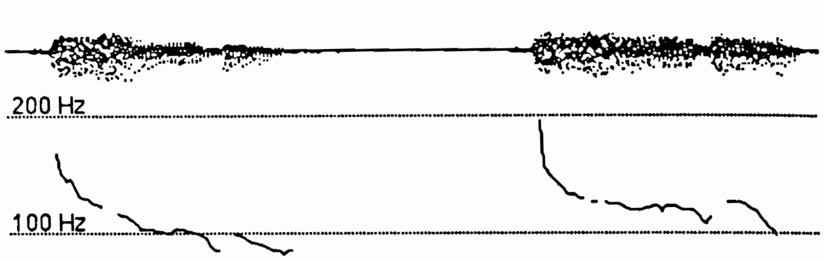


Figure 2. The tonal contrast between /ku'laamba/ 'to lick', and /kulaamba/ 'to be dear', in Sukuma, a Bantu language,spoken in Tanzania.

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. Electroglottograhic data can be recorded in a similar way

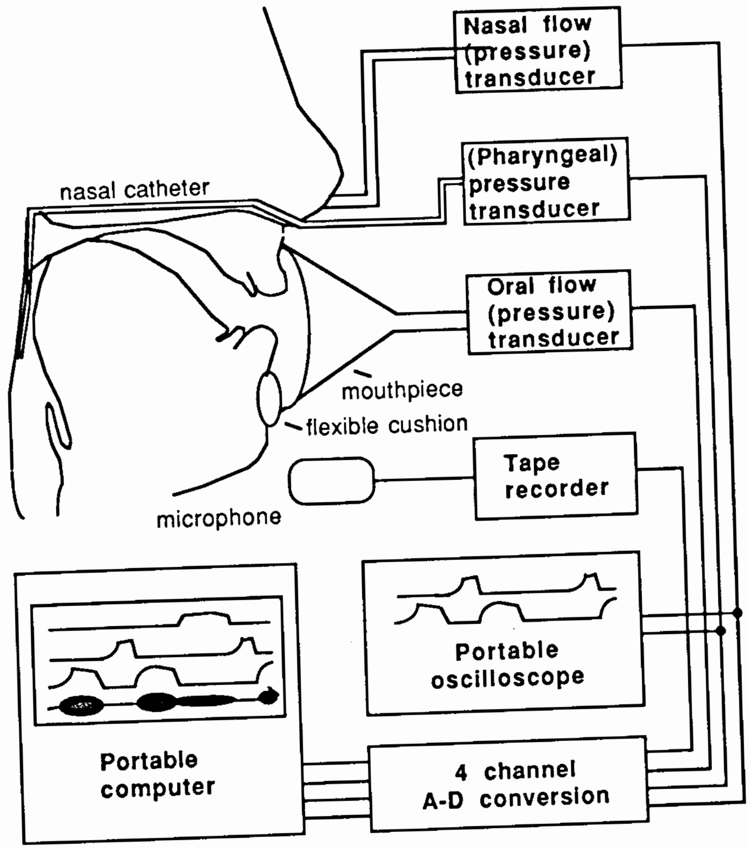


Figure 3. Apparatus for obtaining aerodynamic records in the field.

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 traditional 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, U.S.A.

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/ 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 Sonagraph 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 a 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

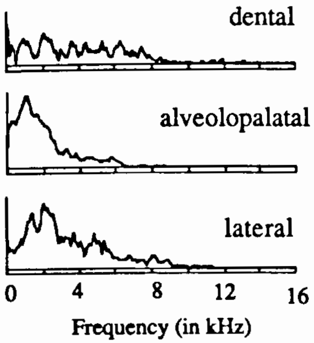


Figure 1: Mean spectra of the dental, alveolopalatal and lateral clicks before the vowels /i,e,a/ for two male Xhosa speakers. Each curve is the mean of six spectra.

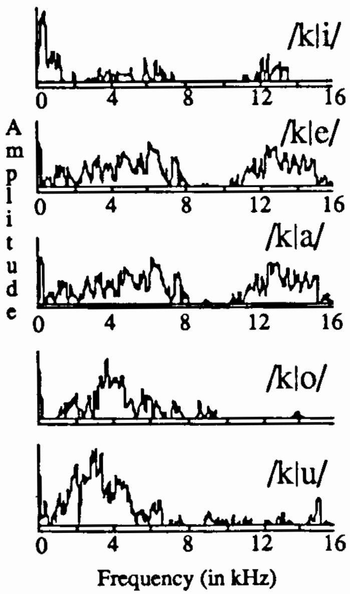


Figure 2: Spectra of voiceless unaspirated dental clicks of one female speaker of Xhosa, before each of the five vowels.

lesser amplitude present up to 16,000 Hz. The amplitude level of the dental clicks is lower than that of the lateral or the alveolopalatal clicks. While all of the dental clicks can be characterized as 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.

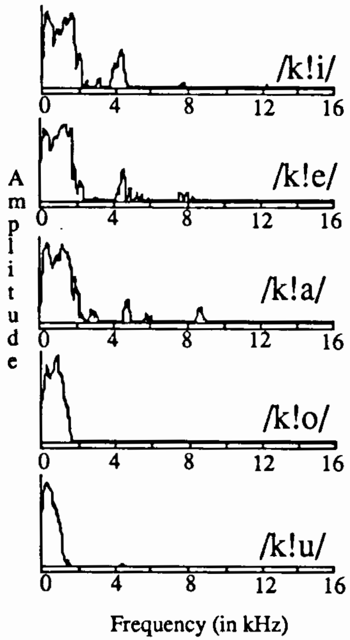


Figure 3: Spectra of voiceless unaspirated alveolopalatal clicks of one female speaker of Xhosa, before each of the five vowels.

For the alveolopalatal clicks, as seen in Figures 1 and 3, 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 characterized as having a compact spectral shape [1].

The effect of a rounded vowel on a preceding click can be seen for the alveolopalatal 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 alveolopalatal 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.

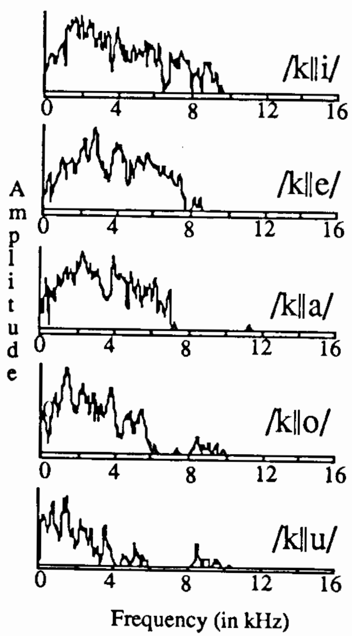


Figure 4: Spectra of voiceless unaspirated lateral clicks of one female speaker of Xhosa, before each of the five vowels.

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 2100-4000 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. This is 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 /o/ and /u/, which is an expected result of anticipation 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 all have 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, o, u/ occuring after dental, lateral and alveolopalatal voiceless unaspirated clicks. The vowels of 7 Xhosa speakers were analyzed. 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.

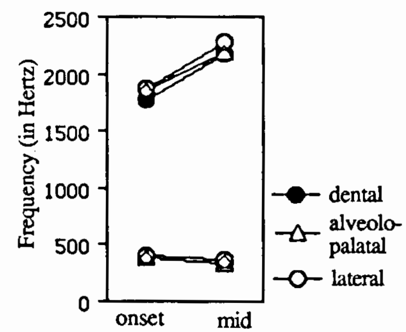


Figure 5: F1 and F2 at onset and middle of /i/, averaged over 7 speakers.

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 onset 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 prodution. 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

[1] BLUMSTEIN S. (1986), "On acoustic invariance in speech", in [4] 178-193.

[2] CLEMENTS G. (1985), "The geometry of phonological features", *Phonology Yearbook* 2. 225-252.

[3] MCCARTHY J. (1989), "Guttural phonology", manuscript. U.Mass Amherst.

[4] PERKELL J. & KLATT D. *Invariance and Variability in Speech Processes*. Laurence Erlbaum. New Jersey.

[5] SANDS, B. (1991), "The acoustic characteristics of Xhosa clicks", UCLA *WPP* 80.

[6] STEVENS, K. (1989), "On the quantal nature of speech", *JOP* . 17.3-45.

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

C. Cucchiarini* & 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 three 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, concentration, auditory acuity 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.

Considering the enormous amount of material in our investigation (8640 transcriptions to be compared thousands of times) it was unthinkable to align transcriptions by hand. A program was therefore developed which makes it possible to automatically 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 cumulative distance between them [1]. String alignment is performed on the basis of distance measures between symbols. If the two transcriptions do not contain the same number of phonetic symbols, null symbols are inserted. On the basis of the distance values, the alignment program determines which symbols are missing in one of the two transcriptions (or have been inserted in the other, depending on the point of view). Owing to space limitations, we cannot go into the difficulties involved in deriving the distance values for transcription evaluation. These difficulties concern not only the choice of the numerical values, but first and foremost the choice of the domain in which speech sounds are to be compared, i.e. perception, acoustics or articulation. Suffice it to say that for want of a better solution we eventually decided to use two matrices, one for vowels and one for consonants, in which sounds are defined by feature values [3]. The features adopted are essentially articulatory. This choice was primarily motivated by the fact that phonetic symbols are defined in terms of articulatory characteristics.

The major differences between our program and that of Picone et al. concern the input matrix:

1. Picone et al. use phoneme distance matrices while our program employs matrices containing feature values. The distances between speech sounds are computed as the program needs them. Although this makes the system slower, it has the important advantage of making it possible to include diacritical marks. Their effect on the different phonetic symbols is computed before determining the distance between two basic symbols.
2. The matrices adopted by Picone et al. contain perceptually based distances, whereas our features are essentially articulatory.
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 per cent 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 this 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 intuitions 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. Morever, it may be argued that if the program provides an undesirable solution it does so on the basis of the knowledge built in it. So, instead of manipulating the outcome one should change the information which led to it. This would imply using the alignment program diagnostically to check whether the distance values are well chosen. For example, if the two following transcriptions are aligned as in 1 while we want the alignment to be as in 2,

```
1  d E n t        2  d E n t
   d E 0 m           d E m 0
```

then it is clear that the distance value between /t/ and /m/ is too small in relation to that between /n/ 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 value 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 null 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 alignment 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 the two strings

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 1. Incorrect alignments

error type	1	2	3	4
cases	7	71	3	6

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 appear counterintuitive it will have to be changed 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, with 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 our case 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 transcription 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 the 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 omitting 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 deletions, 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 and transcription evaluation. Each case will have to be considered separately and the outcome will depend on the purpose of the transcription.

4. REFERENCES

[1] KRUSKAL, J.B. & D. SANKOFF (eds.) (1983), *"Time warps, string edits, and macromolecules: the theory and practice of sequence comparison"*, Reading (Mass.): Addison-Wesley Publishing Company.

[2] PICONE J., K.M. GOUDIE MARSHALL, G.R. DODDINGTON, & W. FISHER (1986), "Automatic text alignment for speech system evaluation", *IEEE Transactions on acoustics, speech, and signal processing, Vol. ASSP-34, No. 4*, 780-784.

[3] VIEREGGE, W.H. & C. CUCCHIARINI (1988), "Evaluating the transcription process", in: Ainsworth, W.A. & J.N. Holmes (eds.) *Proceedings Speech '88.*

(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 Bezooijen* 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 synthesis 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 critisized extensively in the literature, especially with respect to the restrictions imposed on the responses and the limited phonetic contexts in which the target consonants are presented [cf. 3]. The objections raised are particularly serious if 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 comprehending 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 percentages 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 symbol inventory at their disposal to denote subtle and deviant sound characteristics, without any preimposed restrictions. 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, 100 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 several 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 intended 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 the realizations of one target phoneme for 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), phoneme position, and synthesis system.

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 ear-phonetician 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 half of the transcriptions pertaining to the realizations of one 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 overestimation 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 additional information which was considered potentially useful for the improvement of the segmental quality of the synthesis system at hand. The information pertained to voice quality (3 cases), to the undue presence of a final consonant in VCCV 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 proved 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 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, unchecked 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. One needs highly 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

[1] BEZOOIJEN, R. VAN (1990), *Evaluation of speech synthesis for Dutch: comparison of synthesis systems, intelligibility tests, and scaling methods*, SPIN-ASSP Report no. 22, Foundation for Speech Technology, Utrecht.

[2] BEZOOIJEN, R. VAN & POLS, L.C.W. (1987), "Evaluation of two synthesis-by-rule systems for Dutch", *Proc. Eur. Conf. Speech Techn.*, Edinburgh, 1, 183-186.

[3] CARLSON, R. & GRANSTROM, B. (1989), "Evaluation and development of KTH text-to-speech system on the segmental level", *Proc. ESCA Workshop Speech Input/Output Assessm. and Speech Databases*, Noordwijkerhout, 1.3.1-1.3.4.

[4] DUCKWORTH, M., ALLEN, G., HARDCASTLE, W., & BALL, M. (1990), "Extensions to the International Phonetic Alphabet for the transcription of atypical speech", *Clinical linguistics & phonetics*, *4*, 273-280.

[5] GERWEN, R.P.M.W. VAN & VIEREGGE, W.H. (1989), "Evaluation of an automatic text-to-speech conversion system for Spanish", *Proc. Workshop Speech Input/Output Assessm. and Speech Databases*, Noordwijkerhout, 3.5.1-3.5.4.

[6] HOUSE, A.S., WILLIAMS, C.E., HECKER, M.H., & KRYTER, K.D. (1965), "Articulation-testing methods: consonantal differentiation with a closed response set", *JASA*, *37*, 158-166.

[7] POLS, L.C.W., LEFEVRE, J.P., BOXELAAR, G., & SON, N.E. VAN (1987), "Word intelligibility of a rule synthesis system for French", *Proc. Eur. Conf. Speech Techn.*, Edinburgh, 1, 179-182.

CONSONANT CLUSTERS: A COMPARISON BETWEEN WORD INTERNAL AND WORD JUNCTURE

Christine Meunier

Institut de Phonétique, Aix-en-Provence, France

ABSTRACT

We analyze 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 consonant 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 [1]

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

2.2.Consonant clusters classification [2] [3]

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 classes). In these two groups, three types of GC can be deduced from the consonant classification:

Homogeneous GC:
- Ho1 ---> stops + stops
- Ho2 ---> fricatives + fricatives
- Ho3 --->voc.cons.+ voc. cons.

Heterogeneous GC:
- He1 ---> stops + fricatives
- He2 ---> fricatives + vocalic cons.
- He3 ---> stops + vocalic cons.

3-SPEECH MATERIAL

We selected two corpuses. In the first, the Word Corpus (CM, "Corpus Mots" in French), the GC are word internal; word initial for the heterogeneous groups (plat) and medial for the homogeneous ones (obtus). We took into account only the GC from French lexical words. All the words are included in the same sentence: "Ce n'est pas xxxx qu'il faut dire". In the second corpus, the Juncture Corpus (CJ, "Corpus Joncture" 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 comparison:

example:
CC: "ce n'est pas près qu'il faut dire"
CJa: "l'équipe ralentit son allure"
CJb: "ce retard handicape Robespierre"

We recorded two speakers (male) who read the three corpuses 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 GC) 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 GC and for all together.

4.2.Transition phase [2] [3]

An 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 by C1 acoustical characteristics + C2 acoustical characteristics without any other segment.

The Transitory 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 acoustical characteristics of each consonant class:

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

Any possible variations of these simple descriptions (with regard to the phonotypical 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 CJa clusters would be longer than the CJb 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 apparition of pauses between C1 and C2 (evidence of a major boundary).
- the increase of partial and total assimilation numbers in the CJb clusters, and decrease in CJa ones (comparing them to CC clusters).

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

6-RESULTS

6.1.Mean duration :

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

		CM		CJA		CJB	
		M	C	M	C	M	C
ALL GC	C1	105	33	82	34	78	35
	C2	95	34	76	31	67	35
	GC	198	23	159	26	145	30

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), C2 position (C2) and in general (STOP, FRI, VOC) for the three corpuses:

		CM		CJA		CJB	
		M	C	M	C	M	C
STOP	C1	104	32	77	28	77	38
	C2	90	33	60	28	56	23
	STOP	101	32	73	30	73	38
FRI	C1	107	36	96	32	83	29
	C2	88	27	95	25	86	32
	FRI	101	35	96	30	84	30
VOC	C1	106	32	68	39	69	39
	C2	97	35	75	29	65	32
	VOC	99	34	74	31	66	31

We do not notice changes in the three corpuses for stops: stops are always longer in C1 than in C2 position. For fricatives, we see a difference between CC and CJ (a and b): CC fricatives are longer in first than in second position; in CJ (a and b) they 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 CJb (but with a slighter difference), and the opposite for CJa. We must notice the strong stability of CC (whatever the consonant class), and the similarity between CJa and CJb with the exception of vocalic consonants. Consonants seem to be longer in CJa than in CJb.

6.2.Correlations :

Table 3: Correlation matrix of C1, C2 and consonant clusters for the three corpuses in general (number: 92)

		CM			CJA		
	↓	c1	c2	gc	c1	c2	gc
CJA	c1	0.31					
	c2		-0.18				
	gc			0.152			
CJB	c1	0.35			0.317		
	c2		0.255			0.322	
	gc			0.433			0.275

Significant correlations for 0,01 and 0,02 probability : CM/CJb, CJa/CJb, CM/CJa (only for C1).
Not significant: :CM/CJa (for C2,GC).

Table 4: idem table3: Ho1 (number: 16)

		CM			CJA		
	↓	c1	c2	gc	c1	c2	gc
CJA	c1	-0.152					
	c2		0.073				
	gc			0.385			
CJB	c1	0.182			-0.052		
	c2		0.476			0.212	
	gc			0.617			0.252

Significant correlations for 0,01 and 0,02 probability : CM/CJb (GC only).
Not significant: : CM/CJa, CM/CJb (for C1), CJa/CJb.

Table 5: idem table3: Ho3 (number: 20)

		CM			CJA		
	↓	c1	c2	gc	c1	c2	gc
CJA	c1	-0.028					
	c2		-0.133				
	gc			-0.203			
CJB	c1	0.091			0.201		
	c2		0.337			0.341	
	gc			0.25			0.228

Significant correlations for 0,01 and 0,02 probability : none.
Not significant: : all.

Table 6: idem table3: He2 (number: 28)

		CM			CJA		
	↓	c1	c2	gc	c1	c2	gc
CJA	c1	0.425					
	c2		-0.173				
	gc			0.103			
CJB	c1	0.666			0.617		
	c2		0.247			-0.014	
	gc			0.748			0.284

Significant correlations for 0,01 and 0,02 probability : CM/CJa (for C1), CM/CJb (for C1 and GC), CJa/CJb (for C1).
Not significant : CM/CJa (for C2, GC), CM/CJb (for C2, GC), CJa/CJb (for C2).

Table 7:idem table3: He3 (number: 32)

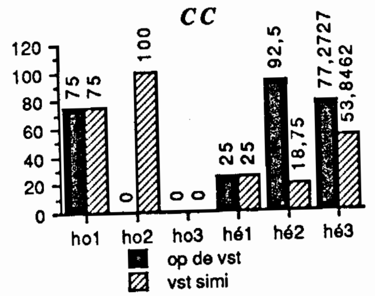
		CM			CJA		
	↓	c1	c2	gc	c1	c2	gc
CJA	c1	0.558					
	c2		0.143				
	gc			0.387			
CJB	c1	0.522			0.159		
	c2		0.288			0.339	
	gc			0.627			0.292

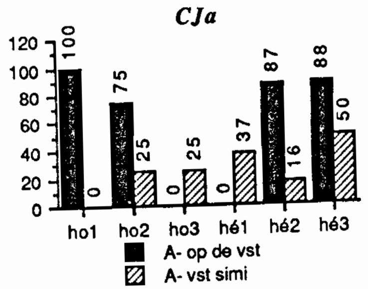
Significant correlations for 0,01 and 0,02 probability : CM/CJa (for C1), CM/CJb (for C1 and GC).
Not significant : CM/CJa (for C2), CM/CJb (for C2), CJa/CJb (for C1and GC).

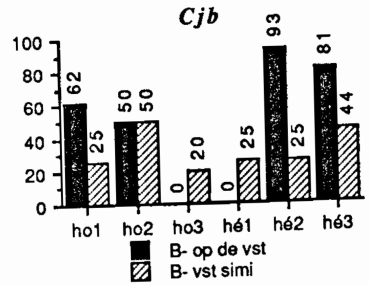
We do not give tables for Ho2 nor He1 because we have not enough values for the results to be relevant. In table 3, all the correlations are significant with the exception of CM/CJa (for C2 and GC); our hypotheses are partially confirmed: there is a better relation between CM/CJb than between CM/CJa. We observe very bad correlations in Ho1 and Ho3 (tables 4 and 5). For He1 and He2 (tables 6 and 7), the correlations are quite similar, with particulary good results for CM/CJb (C1 and C2) and CM/CJa (C1). C2 seems to support variability when we change the context, instead of C1 which is the stable consonant of the cluster in all contexts. But C2, in He2 and He3 is the vocalic consonant and this phoneme seems to be instable in all cases (see table 5).

6.3.Transition Phases

Table 8: Distribution of the Transitory Segments in the six consonant cluster classes for the three corpuses; voicing opposition (opp de vst) and similar voicing (vst simi) inside clusters are separated in each class:







Here our hypotheses are not confirmed: we do not notice a decrease of ST in the CJa realisation, nor an increase in CJb ones. In fact, the tables show stability in the distribution of ST whatever the context. These data confirm the correation results: strong stability for He2, He3; variation for Ho1, Ho3 (Ho2 and He1 values are not sufficient to be significant).

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, completly) devoiced. When the consonants are not in voicing opposition, some ST are also present: it can be an insertion of a vocalic element (particulary in Ho1), or the "consonantification" of the vocalic consonant (/j/ following stops or fricatives). We did not note any pause in CJa context.

7.CONCLUSION

Some of our hypotheses seem to be partially confirmed by the results of the acoustical analysis: CJa clusters tend to be longer than CJb ones; the acoustic organisation of CJb clusters tends to look like CM one, instead of CJa. 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

[1] AUTESSERRE, D., ROSSI, M. (1985), "Proposition pour une segmentation et un étiquetage hiérarchisé. Application à la base de données acoustique du GRECO Communication Parlée", *Actes des 14èmes Journées d'Etudes sur la Parole*, Paris, 147-152.
[2] MEUNIER, C. (1990), "Groupes consonantiques: premier inventaire des réalisations acoustiques des phases de transition", *Actes des 16èmes Journées d'Etudes sur la Parole*, Montréal, 69-73.
[3] MEUNIER, C. (1990), "L'analyse acoustique des groupes consonantiques: deux exemples de groupes hétérogènes" *Proceedings of the LP'90 Conference*, Prague, (in press).
[4] NISHINUMA, Y., et al. (1989), "Duration of Consonant Clusters in French: Automatic Detection Rules" *Proceedings of the European Conference on Speech Communication and Technology*, Paris, 260-264.
[5] ROCHETTE, C. (1973), *Les groupes de consonnes du français*, Québec: Klincksieck.

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 on 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 (de)nasalization. The phonetic as well as forensic implications of the findings are discussed.

1. INTRODUCTION

It is common knowledge among ordinary people as well as phoneticians that the consumption of alcoholic beverages, especially in large quantities, affects the verbal behavior. Yet while the effect of alcohol on certain neurophysiological 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 only very few of them have actually tried to measure the degree of intoxication. Instead, they often had to use 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 5 [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 recent 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 male subjects who were 23 years old on the average (SD = 15 months). The task reported on here was the reading of a phonetically balanced text (The Northwind 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 breathalyser for their breath alcohol level (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 like /ts/, /n/, /l/, /r/ etc. (cf. e.g. [3, 5]), but in analysing our data we found that the segmental perspective was too narrow in order to explain some of the changes observed in the sense that a number of sounds are affected by certain general processes. I will thus try to outline some mechanisms which seem to be affected by alcohol intoxication.

4.1.1. Velar action

The preliminary auditory analysis of the data revealed a marked increase in denasalized articulation of the nasal consonants in intoxicated condition. A systematic evaluation of this phenomenon in relation to the maximum individual intoxication shows that even at very low levels of breath alcohol concentration (i.e below 0.08%), about 30 % of the subjects exhibit an increase in denasalazation of nasals; above 0.08% there is a drastic increase, and above 0.16% all subjects have denasal consonants. In view of this finding we also looked for the complementary effect, namely the nasalization of vowels as compared to the sober condition. Again, the correlation with the degree of alcohol intoxication is obvious, but vowel nasalization sets off at a later stage, i.e. above 0.08% BAL. (Fig. 1) It is important to note that the denasalization of vowels implies the nasalization of vowels, i.e. there is no case of consonant denasalization without vowel nasalization.

We explain these findings by a decrease in velar motility due to impaired motor control. A local effect on the mucosa seems highly improbable since INT-checkups conducted throughout the experiment revealed no effects on the laryngeal or pharyngeal mucosa.

4.1.2. Slurred Articulations

One of the most frequently mentioned effects of alcohol on speech is the so-called slurred or incomplete articulation of segments or clusters which are then "reduced", usually at the expense of the plosive element [2, 4]. The average number of incomplete articulations per person at the maximum individual BAL is increased even at low levels of intoxication as compared to the sober condition; it triples above a BAL of 0.8% and rises again drastically above 0,2%. (Fig. 2) (It has to be emphasized that only the changes compared to the sober condition were taken into account.)

As is shown by an in-depth analysis of the data, the sounds affected by incompleteness are mostly apicoalveolars of different manners of articulation, i.e. plosives, fricatives, nasals, and laterals. This indicates that the motor control of the tip of the tongue, which has to perform the most delicate articulatory movements, is impaired and thus these movements are not carried out completely.

4.1.3. Segment Lengthening

Segment lengthening forms one of the most commonly stated effects of alcohol [2]. The percentage of subjects showing vowel and consonant lengthening rises from 18% (vowels as well as consonants)

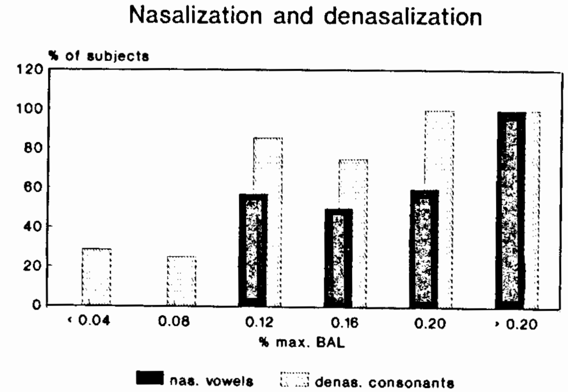


Figure 1

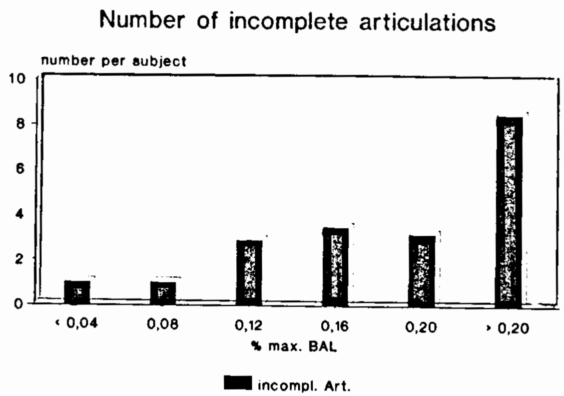


Figure 2

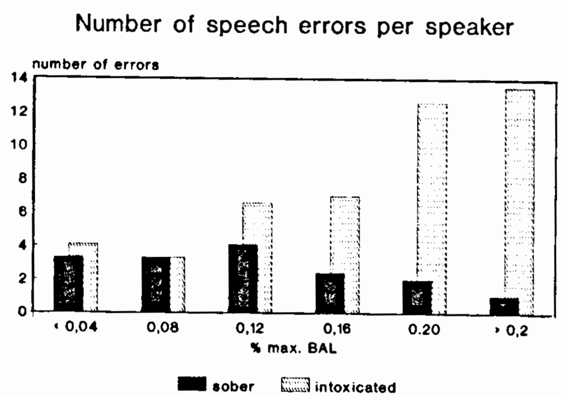


Figure 3

below 0.08% to 50% (consonants) and 81% (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 (slips of the tongue) in the read 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 comparably 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 38% 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 condition. 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 soberness.

6. REFERENCES

[1] ADRIAN, W. (1979), "Bestimmung des Alkoholgehaltes im menschlichen Körper über die Atemluft", *Der Sachbeweis im Strafverfahren, BKA-Vortragsreihe 24*, 99-103.

[2] LESTER, L. / SKOUSEN, R. (1974), "The Phonology of Drunkenness", *Papers from the Parasession on Natural Phonology*, Chicago, 233-239.

[3] PISONI, D. et al. (1986), "Effects of alcohol on the acoustic-phonetic properties of speech". *Alcohol, Accidents, and Injuries, (Society of Automotive Engineers, Warrendale, Pa.) Special Paper P-173*, p. 131-150.

[4] PISONI, D. / MARTIN, Chr. (1989), "Effects of Alcohol on the Acoustic-Phonetic Properties of Speech: Perceptual and Acoustic Analyses", *Alcoholism: Clinical and Experimental Research, 13*, 577-587.

[5] TROJAN, F. /KRYSPIN-EXNER, K. (1968), "The decay of articulation under the influence of alcohol and paraldehyde", *Folia phoniatrica, 20*, 217-238.

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 consonantal 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, voicing and also contrastive 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 more 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 'voicing' in Dutch to the phonetic-acoustic cues 'duration of the vowel' and 'duration of the closure'. As will be evident, short and long vowels differentiate in short vs. long duration while voiced and voiceless plosives are characterized by short vs. long closure duration [4].

Firstly, two basic questions can be formulated as follows: 1) how do young children handle durational values of short vowels as opposed to long vowels 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 voiced 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 coordination 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 was 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)V$CV(C) words with lexical stress upon the first syllable ($=syllable boundary). The intervocalic consonant was either a voiced plosive, that is /b/ or /d/, or a voiceless plosive, that is /p/ or /t/, e.g. the words 'kabel' (cable) vs. 'stapel' (pile). In approximately half of the words the vowel preceding the intervocalic consonant was a phonologically short vowel /ɑ/, /ɔ/, /ɛ/, or /I/, otherwise it was a long vowel /a/, /o/, or /e/. Experimental research with young children imposes several constraints upon the selection of meaningful words to be used: No exact minimal pairs could be found, 11 words with 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 all the words were elicitated 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 elicitated 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 whatsoever.

2.4. Recordings

Recordings of the four-year-old children were made at home with a Tandberg recorder and a microphone Sennheiser MD21HN. The six- and twelve-year-old hildren and the adults were recorded in a laboratory setting with a Revox A77 recorder and an electrolaryngograph to registrate the exact timing of the vocal pulsing. All subjects 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), nor any deletion of intervocalic plosives.

2.5. Measurements

The synchronous audio- and electrolarynx signals were stored digitally on a microVAX II computer and the speech editing system provided visual and auditory information for segmentation. 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 was 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 I. 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.

Short vowels	4	6	adults
/ɑ/	147	125	84
/ɔ/	146	120	81
/ɛ/	152	128	87
/ɪ/	140	119	79
Long vowels			
/a/	239	217	159
/o/	218	184	140
/e/	234	184	140
short	146	121	83
long	233	206	152
ratio	0.63	0.59	0.54

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.

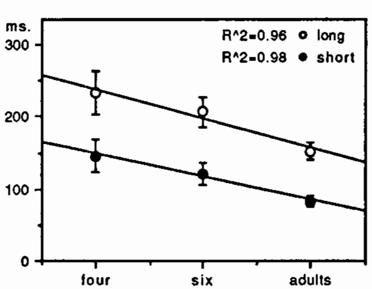


Fig.1 Reduction of short and long vowel duration across groups; regression lines predicting reduction of 'vowel duration' from 'age'.

Regression analysis of the variable 'vowel duration' upon 'age' shows that the proportion of variance 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 plosives /p,t/ and /b,d/ 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 vocalic opposition, the contrast in consonantal closure for /b/ vs. /p/ *increases* with age form 0.58 to 0.66 to 0.72 and for /d/ vs. /t/ from 0.50 to 0.59 to 0.68.

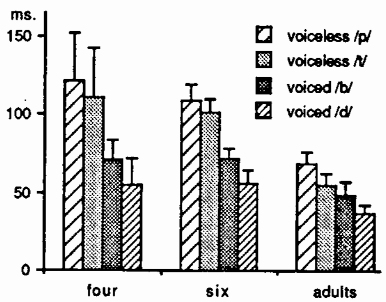


Fig.2. Mean closure durations in ms. for voiced and voiceless plosives in three age groups.

Overall closure duration and relative differences between voiced and voiceless closure durations show no significant differences in speech of four- and six-year old children. Analyses of closure durations between four-year-olds and adults as well as between six-year-olds and adults show significant differences at $p<.01$ or beyond, for both overall duration and relative differences between the voiced and voiceless plosives. Lengthening of the closure durations in speech of young children is certainly commensurate with their slower speaking rate. However, 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 plosive. In Fig.3a, 3b, and 3c behaviour of short and long vowels is plotted for all subjects in the three age groups.

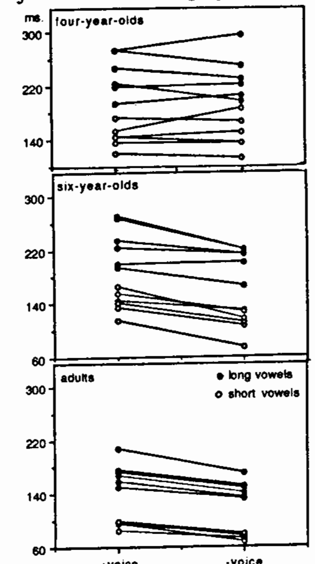


Fig.3a-c. Mean durations (in ms.) for vowels preceding voiced and voiceless consonants in the three age groups Each line represents vowel durations of one subject.

It will be clear that four-year-olds behave very differently from the older children and the 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 voiced 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, indicated that differences in vowel duration preceding a voiceless plosive was not only determined by speaking rate but, again, by some effect due to developmental age.

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 'voicing'. 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 are still in a developmental stage by the age of six. And, contrary to studies using an 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

[1] HAWKINS, S. (1984).'On the development of motor control in speech: Evidence from studies of temporal coordination'. In J. Lass (Ed.), *Speech & Language: Advances in basic research and Practice* (Vol.11). New York: Academic Press, 317-373.
[2] KUIJPERS,C.T.L.(1989).'The voiced-voiceless distinction in speech of four-year-old children.' *Proceedings of the Institute of Phonetic Sciences Amsterdam 13*, 59-75.
[3] LISKER, L. (1978).'Rapid vs. Rabid. A catalogue of acoustic features that may cue the distinction'. *Haskins Laboratory Status Report SR-54*, 127-132.
[4] PORT, R.F. (1981).'Linguistic timing factors in combination'. *J. Ac. Soc. Am. 69(1)*, 262-274.
[5] RAPHAEL, L.J., DORMAN, M.F. & GEFFNER, D. (1980).'Voicing conditioned durational differences in vowels and consonants in speech of three- and four-year-old children'. *J. of Phonetics 8*, 335-341.
[6] SMITH, B.L. (1978).'Temporal aspects of English speech production'. *J. of Phonetics 6*, 37-68.
[7] KENT, R.D. (1980). 'Speech segment durations in sentence recitations by children and adults'. *J. of Phonetics 8*, 157-168.

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 nonmeaningful 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 labials, nasals, and syllabic consonants and a lower proportion of alveolars and supraglottal stops. Group differences increased 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 or later in HI subjects [2] and that the phonetic inventories of NH and HI differed in their size (HI inventories were smaller) and composition [3,4].

Stoel-Gammon's detailed comparison [3] 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 and a lower proportion of stops than the NH group. Since the study focused exclusively on consonantal inventories (i.e., on consonantal types), it provides only a partial picture of the phonetic characteristics of the prelinguistic vocalizations of the two groups.

The present study extends the work Stoel-Gammon [3] by analysing 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 NH 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 3-6. 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 samples. 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 analysed 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 comparisons

Previous studies [2,4] suggested that the vocalizations of HI subjects 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 percentage 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 an even greater magnitude. Table 1 presents a comparison 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 supraglottal 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 supraglottal consonants.

	NH	HI
%Labial (SD)	42.0 (26.5)	71.7 (27.4)
%Alveolar (SD)	34.2 (23.6)	12.1 (15.9)
%Stop (SD)	34.4 (19.3)	14.4 (16.3)
%Nasal (SD)	24.9 (22.8)	50.5 (29.1)
%Syllabic (SD)	22.8 (23.4)	43.2 (28.4)

The comparison of manner features highlights three areas in which the group samples differed markedly: the HI samples contained a much higher proportion of nasal consonants and a much lower proportion of supraglottal stops. In addition, the HI subjects produced proportionally 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 proportional occurrence of features of supraglottal consonants only.

TABLE 2. NH Subjects: Place and manner of supraglottal consonants by age.

age *	Early	Mid	Late
%Lab (SD)	58.9 (27.8)	36.7 (26.6)	32.3 (17.1)
%Alv (SD)	13.9 (14.7)	41.7 (25.2)	46.2 (13.6)
%Stop (SD)	18.0 (15.8)	40.0 (19.4)	43.0 (11.3)
%Syl (SD)	47.3 (24.1)	17.1 (16.1)	6.1 (2.9)

*Early: 5.0-7.3 months (13 samples)
Mid: 8.0-13.6 months (18 samples)
Late: 14.4-18.4 months (13 samples)

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 the Mid age range greatly exceeds the change between the Mid and Late age periods, though the standard deviation declines considerably in the latter period indicating more uniform performance.

For manner of articulation, the mean proportional occurrence of supraglottal stop consonants 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 supraglottal consonants in the Early period to about 6% in the Late period.

Table 3 presents a comparison, based on analysis of supraglottal 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). It is evident from the table 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 decease 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 supraglottal consonants by age.

age *	Early	Mid	Late
%Lab (SD)	37.2 (20.3)	90.8 (7.6)	74.5 (24.3)
%Alv (SD)	23.8 (24.1)	3.8 (2.5)	12.4 (14.1)
%Stop (SD)	20.9 (11.7)	6.9 (6.9)	17.7 (21.2)
%Syl (SD)	50.0 (12.0)	57.5 (33.7)	29.5 (27.5)

*Early: 5.0-12.0 months (7 samples)
Mid: 15.0-21.2 months (9 samples)
Late: 22.7-39.4 months (12 samples)

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 alveolars which decreased from, a mean of 23.8% in the Early period to 3.8% in the Mid period and then increased to 12.4% in the Late period. The proportional occurrence of supraglottal 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 was most similar in the samples from the youngest subjects and became 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 nonmeaningful vocalizations of the NH subjects at this age were not analyzed. It is clear, however, that the U-shaped developmental curves in the HI samples make the productions of the Late period 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 thus their articulation can be seen and imitated by babies who have little or no auditory input; alveolar consonants, by comparison, lack this 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 kinesthetic 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 Institutes of Health: grants R01-HD12695 and P01-NS26521.

REFERENCES

[1] OLLER, D.K. (1986) "Metaphonology and infant vocalization", in B.Lindstrom & R. Zetterstrom (Eds.) *Precursors of early speech*. Basingstoke, Hampshire: Macmillan.
[2] OLLER, D.K. & EILERS, R.E. (1988) "The role of audition in babbling" *Child Development, 59*, 441-449.
[3] STOEL-GAMMON, C. (1988), "Prelinguistic vocalizations of hearing-impaired and normally hearing Subjects: A comparison of consonantal inventories", *Journal of Speech and Hearing Disorders, 53*, 302-315.
[4] STOEL-GAMMON, C. & OTOMO, K. (1986) "Babbling development of hearing-impaired and normally hearing subjects. *Journal of Speech and Hearing Disorders, 51*, 33-41.

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.

<u>Table 1: Details of Children Studied</u>

<u>Sibling 1</u>

Date of Birth:	31.3.77
I.Q.	119
Age of Diagnosis:	3 yrs
Severity of DVD:	Severe
Major difficulties:	Phonology
	Syntax
	Lexicon
	Clumsiness
	Arithmetic
	Auditory Memory

<u>Sibling 2</u>

Date of Birth:	14.1.81
I.Q.	131
Age of Diagnosis:	1 yr 10 m
Severity of DVD:	Severe
Major difficulties:	Phonology
	Syntax
	Lexicon
	Writing
	Spelling

<u>Sibling 3</u>

Date of Birth:	22.3.83
I.Q.	113
Age of Diagnosis:	3 yrs 5 m
Severity of DVD:	Mild
Major difficulties:	Phonology
	Syntax
	Lexicon
	Clumsiness
	Abstract Concepts

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 in 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 monosyllables 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

1. FERGUSON, C.A. and SLOBIN, D.I. (1973), "Studies of child language development", Holt, Rinehart and Winston Inc.
2. MILLOY, N. (1986), "The discovery of a population", Speech Therapy in Practice, 8-10.
3. PRICHARD, C.L., TEKIEL, M.E. and KOZUP, J.M. (1979), "Developmental apraxia: diagnostic considerations", Journal of Communication Disorders, 12, 337-348.
4. TATE, M. (1990), "A study of the speech and language development of three siblings diagnosed as being verbally dyspraxic", Unpublished M.Sc. dissertation, University of Manchester, Faculty of Medicine.

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

C. T. Best[1], A. G. Levitt[2] & G. W. McRoberts[3]

Haskins Laboratories, New Haven CT 06511, USA and
[1]Wesleyan Univ., Middletown CT; [2]Wellesley College,
Wellesley MA ;[3]Stanford 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 sentences. Across age, 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 some 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 niñita más linda!); You are such a great, big boy! (Eres un niño grande!); My beautiful little doll! (Mi muñequita linda!); Who is this little fellow? (Quién es este niñito?); How are you doing today? (Y como estas tú hoy?); And whose sweet baby are you? (De quién es este bebe?). A female speaker of American English, and one of Mexican Spanish, produced multiple IDS tokens 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 questions or the statements 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 1st 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

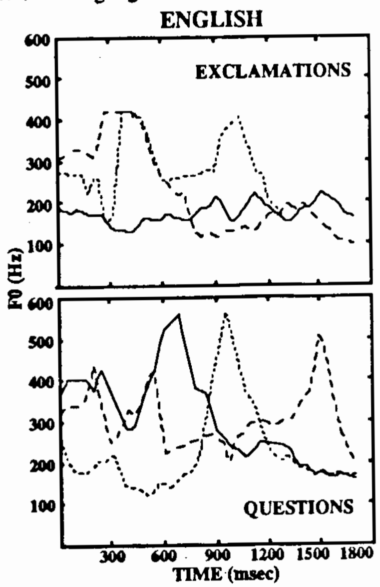


Figure 1. F0 contours (7% smoothing) of English statements (exclamations) and questions.

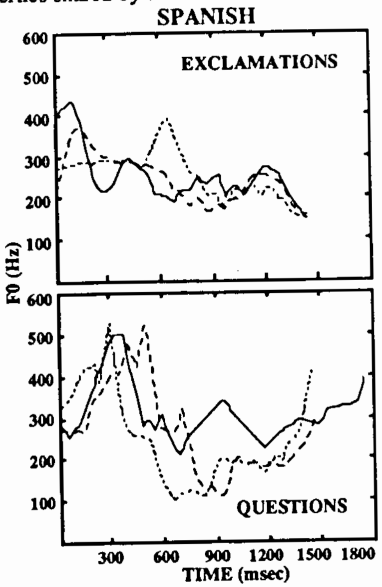


Figure 2. F0 contours (7% smoothing) of Spanish statements (exclamations) and questions.

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 fixations 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 at 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 [$F(1,30) = 10.17$, $p < .001$], 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 patterns.

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 not the arbitrary change, in both English [$F(1,7) = 8.209$, $p < .024$] and Spanish [$F(1,7) = 14.42$, $p < .007$]. 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 between 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 showed 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.

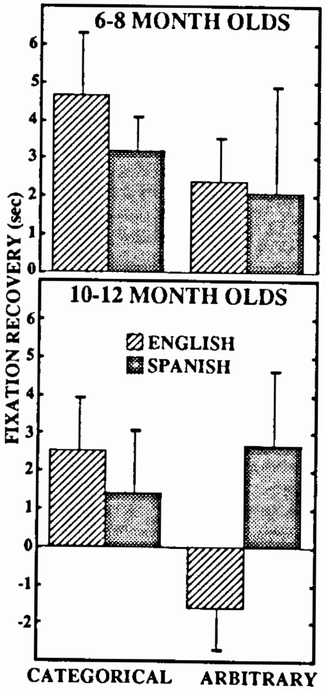


Figure 3. Discrimination in each age and condition, displayed as mean post-shift fixation minus mean pre-shift fixation (bars show s.e.).

This pattern was qualified, however, by the results of separate analyses on each age group. Paradoxically, 10-12 month olds were less able than the younger infants to recognize and discriminate the prosodic categories than were the 6-8 month olds. Nor did the performance of either group reflect language-specific reorganization in perception of prosodic contrasts. The younger infants discriminated the categorical change in both languages, but the older infants' 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 redundant, highly-emphasized references to objects and 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 less attentive to prosodic properties, and more focused on segmental and/or lexical information, than are younger infants.

This study provided little evidence for earlier attunement to native prosodic contrasts than to segmental contrasts. On the contrary, the 10-12 month reorganization in perception of non-native segmental contrasts does not appear to be preceded or even paralleled by analogous reorganization in the perception of this linguistic prosodic contrast.

5. ACKNOWLEDGMENT.

Supported by NICHD grant HD-01994 to Haskins Laboratories and NIDCD grant DC-00403 to the first author.

6. REFERENCES

[1] BAHRICK, L., & PICKENS, J. (1983), "Classification of bimodal English and Spanish passages by infants," *Infant Beh. Dev.*, *11*, 277-296.

[2] BEST, C. (in press), "The emergence of language-specific phonemic influences in infant speech perception," in H. Nusbaum & J. Goodman (eds.) *The transition from speech sounds to spoken words:,"* Cambridge MA: MIT.

[3] BEST, C., McROBERTS, G., & SITHOLE, N. (1988), "The phonological basis of perceptual loss for non-native contrasts: Maintenance of discrimination among Zulu clicks by English-speaking adults and infants," *J. Exp. Psy: HPP*, *14*, 345-360.

[4] DECASPER, A., & FIFER, W. (1980), "Of human bonding: Human newborns prefer their mothers' voices," *Science*, *208*, 1174-1176.

[5] DECASPER, A., & SPENCE, J. (1986), "Prenatal maternal speech influences newborns' perception of speech sounds," *Inf. Beh. & Dev.*, *9*, 133-150.

[6] FERNALD, A., & KUHL, P. (1987), "Acoustic determinants of infant preference for motherese speech," *Inf. Beh. & Dev.*, *10*, 279-293.

[4] FERNALD, A., TAESCHNER, T., DUNN, J., PAPOUSEK, M., BOYSSON-BARDIES, B & FUKUI, I (1990). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants," *J. Child Lang.*

[5] HIRSCH-PASEK, K., KEMLER NELSON, D., JUSCZYK, P., WRIGHT CASSIDY, K., DRUSS, & KENNEDY, L. (1987), "Clauses are perceptual units for young infants," *Cogn.*, *26*, 269-286.

[6] JUSCZYK, P. (1989), "Perception of cues to clausal units in native and non-native languages," Presented at *Soc. Res. Child Dev.*, Kansas City, April.

[7] KAPLAN, E., & KAPLAN, G. (1970), "Is there any such thing as a prelinguistic child?" in J. Eliot (ed.) *"Human development and cognitive processes."* New York: Holt, Rhinehart, & Winston

[8] KEMLER NELSON, D. (1989), "Developmental trends in infants' sensitivity to prosodic cues correlated with linguistic units," Presented at meeting of *Soc. Res. Child Dev.*, Kansas City, April.

[9] LEWIS, M. (1936), *"Infant speech: A study of the beginnings of language."* New York: Harcourt, Brace, Jovanovich.

[10] LISKER, L., & ABRAMSON, A. (1970), "The voicing dimension: Some experiments on comparative phonetics," *Proc. 6th Int. Cong. Phon. Sci.*, Prague: Academia.

[11] MACKAIN, K., BEST, C., & STRANGE W. (1981), "Categorical perception of English /r/ and /l/ by Japanese bilinguals," *Appl Psycholing.*, *2*, 369-390.

[12] MEHLER, J., JUSCZYK, P., LAMBERTZ, G., HALSTED, N., BERTONCINI, J., & AMIEL-TISON, C. (1988), "A precursor of language acquisition in young infants," *Cognition*, *29*, 143-178.

[13] SNOW, C. (1977), "The development of conversation between mothers and babies," *J. Child Lang.*, *4*, 1-22.

[14] WERKER, J., & LALONDE, C. (1988), "Cross-language speech perception: Initial capabilities and developmental change, *Dev. Psy.*, *24*, 672-683.

[15] WERKER, J., & TEES, R. (1984), "Phonemic and phonetic factors in adult cross-language speech perception," *JASA*, *75*, 1866-1878.

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 contrast 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 underived, 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. However, recent studies have found evidence for a continuous line of development from prelinguistic mouthings through babble to 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 routines 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 coproduction (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 coproduction in CVCV contexts, similar to previous studies [cf. 4, 5]. The utterance types were three nonsense disyllables: ['sasa], ['sisi], and ['susu]. The vowels, [a, i, u], were chosen because they occupy extreme points in the vowel space so that if 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 child 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 vowl) and (e) the midpoint between (b) and (d) (1/4 vowl). Estimates of the center frequencies of the second formants were made at these five locations from Discrete Fourier Transform spectra, computed with a 25.6 msec. 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, 77% at 3/4 fric; 76% of the children's data permitted estimates at 1/2 fric, 85% at 3/4 fric. Estimates 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 a given formant 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 adults, 32-month-olds, and 22-month-olds respectively. In Figure 1a (adults) the F2 measurements at 1/2 fric are virtually the same before all three vowels. At 3/4 fric a front back distinction begins to appear with differences of about 300 Hz between F2 values before [i] and the back vowels. Finally, at 1/2 vowl a vowel space has emerged in which [u] has clearly 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 vowl 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).

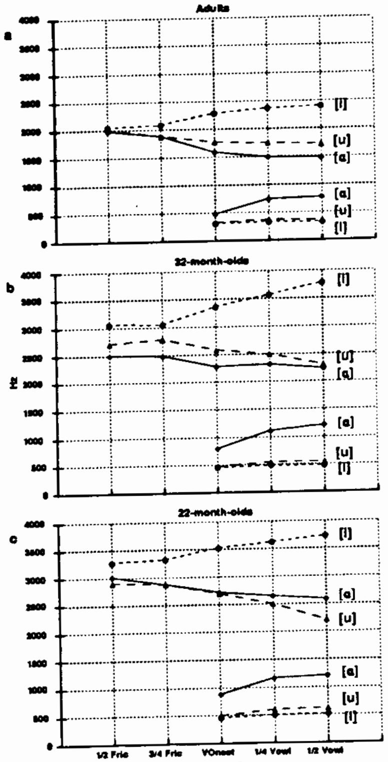


Figure 1a, b, and c. [sV] formant paths for children and adults.

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 before [u] and [a], and the final values of F2 at 1/2 vowl for [u] and [a] reverse the pattern observed in the adults. Both the latter effects arise from an overall higher formant path 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 [i] 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 < .025, one-tailed t-test.

Measurement Point	*i/u*	*i/a*
1/2 Fricative		
22-month-olds	1.15*	1.12*
32-month-olds	1.15	1.24*
Adults	1.00	1.04
3/4 Fricative		
22-month-olds	1.19*	1.21*
32-month-olds	1.13	1.24*
Adults	1.12*	1.13*

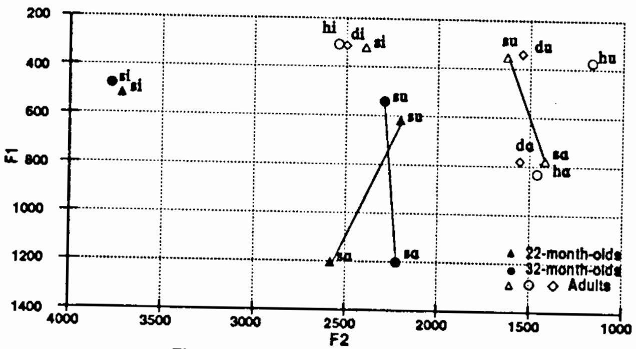


Figure 2. Group vowel plots for selected tokens.

3.2 Gestural Overlap in the Vowel

As noted above, the relative positions of [u] and [a] at 1/2 vowl 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 constriction location for [u] than [a]. For the 32-month-olds, F2s 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 u-a 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 sibilants? To answer this question, more data were collected from the adult subjects. In addition to repeating the original fricative stimulus items ['sisi], ['sasa], and ['susu] 6 times each, subjects also produced 6 repetitions each of ['didi], ['dada], ['dudu], ['hihi], ['haha], and ['huhu]. The means for the first vowels in these contexts are also given in Figure 2. Orthogonal comparisons reveal that [du] and [su] do not significantly differ from each other, [F=1.409, p >.2415], but do differ significantly from [hu], [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. ['seysa] instead of ['sasa]. 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 front-back 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 32 months between fricative F2s for [u] and [a] at 1/2 fric than between F2s for [u] and [a] themselves at 1/2 vowl; (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's speech by differentiation of the closing and opening gestures of the canonical syllable [cf. 7]. Such an account obviates the necessity for positing phonemes, or their featural descriptors, as underived, phonological primitives.

Acknowledgement: Preparation of this paper was supported in part by NIH Grants HD-01994 and DC-00403 to Haskins Laboratories, 270 Crown St., New Haven, CT, USA.

5. REFERENCES

[1] Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology 6*, 151-201.

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

[3] Menn, L. (1986). Language acquisition, aphasia and phonotactic universals. In Eckman, F.R., Moravcsik, E.A. & J.R. Wirth (Eds.) *Markedness*. (pp. 241-255). New York: Plenum Press.

[4] Nittrouer, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research, 32*, 120-132.

[5] Soli, S. D. (1981). Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *Journal of the Acoustical Society of America, 70*, 976-984.

[6] Studdert-Kennedy, M. (1987). The phoneme as a perceptuomotor structure. In A. Allport, D. MacKay, W. Prinz, & E. Scheerer (Eds.), *Language perception and production* (pp. 67- 84). London: Academic Press.

[7] Studdert-Kennedy, M. & Goodell E.W. (submitted). A child's entry into the lexicon: Evidence for a gestural model of early child phonology.

RHYTHMIC PHENOMENA IN A CHILD'S BABBLING AND ONE-WORD SENTENCES

M. Kohno and T. Tsushima

Kobe City University of Foreign Studies

Kobe, Japan.

ABSTRACT

A baby's babbling and one-word sentences, in total 2848 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 mode of long-short timing alternation, while simple one-syllable-repetition-type babbling (e.g. tatatata--) reveals no distinguishable rhythmic pattern, 2) acquisition of isochronism of morae is far later than that of syllables, 3) interstress intervals between syllables in both babbling and one-word utterances become 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 3 to April 9, 1988, which corresponded to the period of her 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 way that Ingram and Menyuk 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 (ILS) run by Micro PDP 11/73, AD Conversion :Das-Box, but Yokokawa Electro-Oscilograph, type 2901, connected with Amplifier 3125, was also used supplementarily.

2. ABOUT BABBLINGS

2.1. Syllabic Constitution

Intervals among voice-onset points of syllables (inter-stress intervals, ISI henceforth), especially of syllables which have plosive-like sounds as consonant partners of CV constructions, were instrumentarily measured. The numbers of utterances thus measured were 130 groups, 245 successions 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 types of 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 types 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.

Table 1 Syllable constitution of Babbling

		7 syl.	6 syl.	5 syl.	4 syl.	Example
a	2 syllable alternate repetition	41	4	23	14	[bakobako-bakoba:]
b	2 syl. alternate repetition in part	7	0	16	2	[bagodago, bagodagi]
c	mono-syllable simple repetition	6	3	12	8	[tatatata-ta:]
d	mono-syl. simple repetition in part	2	0	4	4	[tatatate-to:]
e	no repetition	5	2	8	1	[pikoideo]
	sum.	61	9	63	29	

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 2 ISIs among syllables in babbling

		repetition of one syllable	repetition of two syllables
4 syllables	AVERAGE	313. 9	335. 8
	S.D.	63. 3	118. 3
	N	7	14
	CORRELATION	r=-0. 23	r=-0. 58
5 syllables	AVERAGE	400. 9	417. 8
	S.D.	113. 1	82. 2
	N	13	23
	CORRELATION	r=-0. 05	r=-0. 25
6 syllables	AVERAGE	461. 3	414. 3
	S.D.	109. 1	259. 9
	N	3	6
	CORRELATION	r= 0. 40	r=-0. 05
7 syllables	AVERAGE	498. 4	416. 5
	S.D.	174. 1	91. 3
	N	6	41
	AVERAGE	r=-0. 46	r=-0. 37

In Table 2 which shows means of ISIs (ms) among syllables, S.D. and auto-correlations among 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 except those in the cases of five syllable babblings, 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 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 ones, all occur in the late period, especially in Session IV, except that one example occurred in Session II. 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.

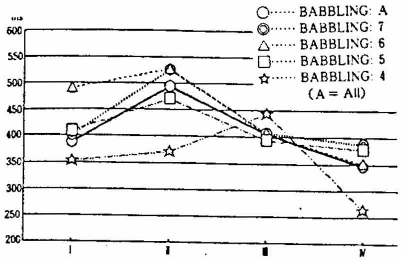


Fig. 1. Chronological Changes of ISIs

Observing chronological change of the ISIs among syllables except the last ones, we can see the ISIs in the early period are longer than the ones in the late period and statistical significance at the level of $p<0.05$ was detected between them. If we examine this phenomenon more precisely according to the units of session, they become longer ($\bar{x}$=387.7 ms to 493ms) during the time lapse from Session I to Session II, immediately before the term in which the subject's vocabulary sharply increases, and then the ISIs become shorter again to Session IV. Fig.1 illustrates this phenomenon graphically.

The vowel lengths were also measured and the fact was found that they scarcely change chronologically from the early to the late periods and so does the S.D.. As shown in Fig.1, on the other hand, the ISIs among syllables vary very much in the passage of time. These facts suggest us that consonants, not vowels, cause chronological changes of ISI such as shown in Fig.1 --in other words, the subject concentrates her attention on the articulations of consonants very much before she can produce plenty of single word utterances in session III, and this results in the expansion of ISIs at session II.

Number of syllables of one succession was counted 2 (minimum) to 7 (maximum) throughout all the recorded data of 455 babbling groups. 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(7±2) of perceptual sense unit (cf. [4]), that is, the chunking unit of utterance which is holistically perceived with its meaning and stored in echoic memory in an unprocessed form in the process of listening comprehension[2].

Table 3 Recorded and Analyzed Data
(single word utterances)

	All	Early	Late	Subject's Renditions
				(broadly transcribed)
Recorded Tokens	2 3 9 4	8 8 8	1 5 0 6	
Analyzed Tokens	2 6 8	1 0 1	1 6 7	
2 − 2 moare				
aka	5 7	1 6	4 1	[aka], [aga]
kabu	1 8	0	1 8	[kobu], [gabu]
choko	4	0	4	[koto]
kiku	2	0	2	[geku]
choki	1	0	1	[didi]
heso	1	0	1	[edo]
jyajya	1	1	0	[dada]
kore	1	1	0	[kode]
Sum	8 5	1 8	6 7	
3 − 3 morae				
akete	4 2	2 2	2 0	[akete], [agete]
okashi	7	0	7	[okati]
kinoko	2	0	2	[gigogo]
osoto	2	0	2	[ototo]
poteto	2	0	2	[poketo]
asoko	1	0	1	[akoko]
Sum	5 6	2 2	3 4	
2 − 3 morae				
dakko	4 6	2 3	2 3	[ga˅ ko], [gago]
totte	3 0	1 5	1 5	[to˅ te], [do˅ te]
cho: dai	4 7	2 3	2 4	[to: de], [do: de]
denwa	4	0	4	[de: ba], [de: ba]
Sum	1 2 7	6 1	6 6	

3. ABOUT SINGLE WORD UTTERANCES

Table 3 shows the number of the recorded data of single word utterances and the contents of acoustically analyzed data.

Table 4 shows syllable intervals (ISIs) in 2 syllables and 2 morae words (2-2 words, henceforth) and in 2-3 words. Just like the case of babbling, the ISIs in one word sentences are longer in the early period than in the late one ($p<0.01$), and the general means of ISIs through the both periods are again similar with the ones of babbling (300-400ms). It might be rightly said,

therefore, that the same timing control mechanism is working in babbling and one-word utterances.

Table 4 Chronological Changes of
ISIs (single word utterances)

	2-2 morae	3-3 morae	AVERAGE
AVERAGE: A (ms)	3 3 4. 0	3 2 8. 7	3 3 1. 9
AVERAGE: E (ms)	3 5 9. 4	3 7 8. 5	3 6 9. 8
AVERAGE: L (ms)	3 2 7. 2	2 9 6. 5	3 1 6. 8
S. D.: All	7 1. 2	6 3. 6	6 8. 1
S. D.: Early	8 2. 8	5 0. 7	6 6. 8
S. D.: Late	6 6. 9	4 8. 8	6 2. 8
N: All	8 5	5 6	1 4 1
N: Early	1 8	2 2	4 0
N: Late	6 7	3 4	1 0 1

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,--short, long, short, shortest intervals in Sessions I, II, III and IV, respectively. More precise observation however, makes it clear that 2-3 words shape this pattern most remarkably,--'dakko' (hold me in your arms), for instance, takes longer time for the transit from 'da' to 'ko' than 'cho' to 'ko'in 'choko' (chocolate). This suggests that the infant already notices the existence of mora in Japanese timing system, but this timing is soon vanished in Sessions III and IV. Has the subject, in the world, mastered the Japanese mora system? In order to make it clear, we carried on the following investigation.

As shown in Table 5, the ISIs in 2-3 words were significantly longer than the ones in 2-2 or 3-3 words ($p<0.01$) not only in the early period but also in the late period (Table 8). Throughout all the periods from the early to the late periods, the means of ISIs in 2-3 words were 457.3ms and the ones in 2-2 words were 334ms and statistical significnce 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' (stump), 'aka'

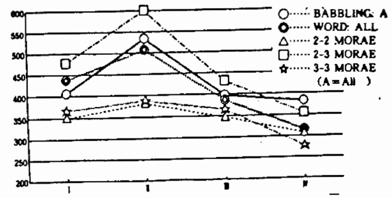


Fig. 4 Chronological Changes of ISIs
(single word utterances)

Table 5 Comparison of Mora Lengths

Early

	2-2 morae	3-3 morae	2-3 morae
2 − 2		$t=0.83$	$t=5.22$**
3 − 3			$t=5.45$**

Late

	2-2 morae	3-3 morae	2-3 morae
2 − 2		$t=2.59$	$t=7.59$**
3 − 3			$t=5.30$**

** $p<.01$

(red) (2-2 words), 'dakko', 'totte' (Take it) and 'ke:ki' (cake) (2-3 words) in citation form and in a carrier sentence 'watashiwa ---- to iimasu' (I say ----.), and after having recorded these utterances, the authors instrumentally measured the ISIs between the first and second syllables of these words by the use of ILS. The results were 158.3 and 184.2ms in 'kabu' and 'aka' (2-2 words) respectively, but in the case of 2-3 words, 'dakko' 'totte' and 'ke:ki', the ISIs were 387.5, 382.0 and 363.3ms, that is, the ratio of ISIs was roughly 1:2 between 2-2 and 2-3 words. The infant's ISIs in these words ($\bar{x}$ throughout the whole period) were 'aka': 341.3ms, 'kabu': 307.2ms, 'dakko': 396.6ms and 'totte': 407.8ms, and the ratio between 2-2 words and 2-3 words were therefore 1:1.2-1:1.3. Even in the Session II, in which ISIs widened most, the ratio is only 1:1.6. All the above data show us that the infant, although she perhaps knows the existence of mora, cannot produce isochronism per mora. As for the isochronism per syllable, on the other hand, she has already mastered: in 'akete' for example, the ISIs between 'a' and 'ke', and 'ke' and 'te' were 324 and 367ms respectively, and their ratio is 1:1.1. Mora, isochronous timing system peculiar to Japanese, is so difficult to be mastered in comparison with the one among syllables.

L' INTONATION DE QUESTION DANS LE LANGAGE EMERGENT

Gabrielle KONOPCZYNSKI

Laboratoire de Phonétique, Besançon, France

The child's questioning process, which is one of the basic functions of langage, is examined from its very beginning (09;) to 24;. It appears much earlier than generally suggested by the literature. Linguistic and acoustic parameters (Fo, form of melodic curve, duration, intensity) are investigated in 6 children. It appears that all the interrogative sentences lacking a question word are over-marked in all their parameters.

Le processus de questionnement chez l'enfant, qui a donné lieu, il y a quelques décennies, à une abondante littérature, semble moins à la mode actuellement. Pourtant, le questionnement constitue non seulement une des fonctions de base du langage (BUHLER) mais il est en outre lié à la fois à la progressive socialisation de l'enfant (STERN, LEWIS) et à son développement cognitif (PIAGET, INGRAM, VIGOTSKY). Les problèmes les plus souvent évoqués furent la fonction du questionnement et son origine, tant au plan phylogénétique qu'ontogénétique. La discussion réside dans le fait que les disciples de PIAGET estiment la mise en place des capacités cognitives comme un préalable au développement du langage alors que pour les disciples de VIGOTSKY le langage et surtout la question seraient des outils qui permettent l'établissement des structures cognitives. Nous laisserons les philosophes du langage débattre de ce problème pour remarquer, sur un plan plus concret, que dans les études sur le développement cognitif, l'accent est mis sur le fait que le questionnement de l'adulte vers l'enfant, est quantitativement beaucoup plus fréquent que dans les interrelations adulte-adulte.Ce questionnement de l'adulte diminue avec son augmentation parallèle chez l'enfant. Enfin, le problème de la question a été surtout étudié dans le cadre de l'acquisition de la syntaxe en liaison avec l'ordre des mots et l'apparition des marqueurs syntaxiques (BELLUGI, SLOBIN, BROWN, FERGUSON). Quel que soit le point de vue adopté, l'apparition du questionnement est toujours signalée au plus tôt avec l'émergence des premiers mots, vers 15 mois. Son éventuelle existence antérieure est totalement passée sous silence. Il est vrai que les auteurs recherchent tous le questionnement au niveau du seul lexique. HERMANN [3] qui, dans un ouvrage sur la question dans les langues du monde, consacre un chapitre à l'ontogénèse et à la phylogénèse du questionnement, est un des rares à signaler que l'enfant sait interroger avec la mélodie, mais il situe également cette capacité à deux ans seulement. FONAGY [2] enfin dit explicitement que l'intonation interrogative est apparue tardivement (1;9) chez ses deux enfants parlant le hongrois alors que l'appel existait dès 12; sans que le langage articulé soit en place.

L'aspect perceptuel a été encore plus négligé que la production. Des recherches récentes commencent enfin à s'intéresser à la perception des éléments prosodiques par l'enfant. Les travaux en cours aux Laboratoires Haskins (BEST & al) sembleraient montrer que la discrimination entre interrogatives et assertives serait précoce (06;) et existerait antérieurement à la distinction de contrastes segmentaux.

Nos données qui consistent en une étude très fine des émissions de 6 sujets entre 9 et 24 mois, ainsi que celles de KASSAI [4,5] montrent une acquisition précoce du questionnement chez l'enfant. Les interrogations existent dès 9/10 mois; encore rares (mais la situation choisie

peut en être la cause), elles augmentent en nombre entre 12 et 16 mois; cette croissance devient encore plus rapide avec l'apparition du lexique. Cette augmentation rapide des questions entre un et deux ans s'explique par la mise en place du langage référentiel.

1. METHODE

Le statut de question des énoncés enfantins a été défini par une double procédure: analyse de la situation d'énonciation et analyse auditive du corpus enregistré. Dans l'analyse de la situation d'énonciation, est défini comme question un énoncé reconnu comme tel par l'entourage du bébé. Notons que l'interactant dispose de divers indices, notamment mimo-gestuels, permettant l'attribution d'un sens global ou d'une modalité à un énoncé encore inarticulé; à ces énoncés interrogatifs du bébé, l'adulte réagit généralement soit par une reprise articulée et présentant une extension de la question, soit par une réponse. Pour l'analyse auditive, nous avons travaillé sur la seule bande sonore des émissions des bébés. Il était demandé à 12 auditeurs formés en linguistique mais non informés du sujet exact de la recherche, de catégoriser, lorsque cela leur semblait possible, les énoncés en diverses modalités. C'est ainsi qu'ont été définies, sur la base des seules informations contenues dans l'onde sonore, des modalités telles que appels, énoncés énonciatifs, phatiques, exclamatifs, impératifs....Ces énoncés ont ensuite été soumis à une analyse acoustique destinée à découvrir quels sont les traits acoustiques (Fo, forme de la courbe mélodique, Durée, Intensité,) qui induisent une telle interprétation de la part de l'adulte.

Dans notre étude, nous séparons le questionnement émis en proto-langage durant la période charnière (9-12;) de celui fait à l'aide du premier langage articulé entre un et deux ans, qui est caractérisé par l'émergence progressive du lexique; nous n'étudierons ici que les structures interrogatives sans mot-outil de questionnement; le questionnement avec mots-outils, qui apparaît vers 20 mois, est réservé pour un travail ultérieur.

2. RESULTATS.

2. 1. Période charnière (9-11;).

L'analyse acoustique n'a pu être réalisée que sur un seul sujet féminin, pour des raisons techniques (mauvais rapport S/B des autres enregistrements). Les interactions avec l'adulte étaient rares, en raison de la situation expérimentale choisie, qui consistait à laisser l'enfant jouer seul dans sa chambre et à n'intervenir qu'en cas de demande urgente de sa part. De ce fait, peu d' énoncés interrogatifs ont été produits. Mais l'observation, sans enregistrements exploitables acoustiquement, de plusieurs bébés dans des situations d'interaction, a permis de constater que ce type d'énoncés est relativement fréquent dès 9/10 mois, moins cependant que les énoncés de type phatique ou énonciatif par exemple.

Dix énoncés ont pu être analysés acoustiquement. Ils présentent des constantes certaines : énoncés brefs (2-3 syllabes); Fo moyen : 450 Hz; contour toujours ascendant dans la zone 3-4, c'est-à-dire entre 420 et 600 Hz (pour le problème de la détermination d'une grille de niveaux pour voix enfantines, cf. [6]) avec une dynamique d'une octave environ. Leur Fo initial est toujours supérieur au Fo-usuel , puisque ces énoncés débutent à 400 Hz ou au-dessus (M: 428 Hz) , alors que le Fo-u de ce sujet est de 340Hz [6]. Leur intensité est forte (supérieure à 30 dB), néanmoins inférieure à celle des énoncés phatiques avec lesquels la catégorie des interrogatives partage la zone de tessiture employée (3-4) et souvent, mais pas exclusivement, la forme ascendante de la courbe mélodique. L'intensité des interrogatives forme un pic, avec montée et chute rapide, alors que dans les phatiques, elle est généralement croissante.

On savait depuis longtemps que l'enfant sait questionner, dans la plupart des langues du monde, avec la seule intonation, les mots-outils interrogatifs étant acquis plus tardivement. Mais l'on pensait que le questionnement ne pouvait apparaître qu'avec les premiers mots, comme nous l'avons rappelé ci-dessus. Nos données, confirmées par celles de quelques rares autres travaux [6] montrent donc qu'il n'en est rien: les interrogations, peu nombreuses certes, existent 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 qu'aux mois 9; et 10; car situées plus haut dans la tessiture (niveau 4-5,jusque 900Hz, cf. [6]). Elles dépassent en hauteur les appels, qui ont une courbe ascendante analogue, mais leur intensité est plus faible : la courbe d'intensité, qui est toujours parallèle au Fo, sauf une rapide chute finale, dépasse rarement 40 dB. Beaucoup de ces questions sont monosyllabiques, de type [æ?]), de durée brève (M.= 255ms., extrêmes 140-450ms.), alors que les vocoïdes à fonction non communicative du Jasis sont toujours très longs (M= 967ms, extrêmes jusqu'à 8530ms.)

2. 2.2. Les questions articulées sans mots outils.

Une distinction s'impose à l'intérieur de cette catégorie entre questions marquées uniquement par la mélodie et celles marquées par un mot-outil. Les premières sont les seules attestées jusque vers 20 mois, âge auquel commencent à apparaître les mots interrogatifs qui sont dans l'ordre : [kesəse] et ses diverses formes, [u] = *où* , [komã] = *comment* (22; un seul exemple chez un sujet). Les questions sans mot outil sont formées essentiellement d'énoncés bi- ou trisyllabiques, représentant des objets ou des actions dont l'enfant cherche à connaître le nom. La forme attestée est soit la forme simple, soit le mot précédé de [e], de [æ] ou de [se] formant un ensemble dont le statut est difficile à déterminer: encore mono-mot ou déjà combinaison de deux éléments? Souvent en effet 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 tessiture 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 phatiques est quelquefois plus prononcé. L'apparition du mot permet de réduire la redondance : le Fo baisse et la zone vocale utilisée se restreint. Très souvent, dès qu'il a obtenu une réponse de l'adulte, l'enfant oppose à la forme interrogative la forme énonciative ou impérative du même mot. On a par exemple:

ENFANT	ADULTE
19; - c'est chien? (/)	oui, c'est un gros chien.
- chien (\)	
22; Sophie debout dans sa baignoire, regardant sa mère:	
- assis? (/)	oui, assieds-toi.
- 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. Dans le premier cas c'est l'adulte qui initialise le dialogue en disant un mot quelconque, généralement désignation d'un objet (*c'est un...*) ou d'une action (*on va ....*). L'enfant, qui paraît 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'appropriation du lexique. Ces diverses formes ascendantes sont beaucoup plus marquées que les questions habituelles ; c'est pourquoi il nous paraît difficile de les appeler "questions-échos" comme le proposent BOYSSON- BARDIES & al. [1] dans leur étude du babillage tardif. 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 pente douce comme s'il y avait hésitation. 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 apparemment celle d'une interrelation : regarder avec l'enfant un

catalogue. A 14 mois, l'adulte mène la danse; la participation verbale de l'enfant est essentiellement de type mélodique ou onomatopéique. Vers 18-20 mois, il en va de même, mais l'enfant répète les mots en se servant de la stratégie décrite ci-dessus. Enfin, il finit par jouer lui-même au jeu des questions-réponses : montrant un objet, il dit son nom avec intonation ascendante, et enchaîne immédiatement la réponse avec intonation descendante, sans attendre d'acquiescement de la part de l'adulte; ce dernier ne lui sert pas d'interlocuteur, mais simplement d'oreille réceptrice. Il semblerait que ce soit là une fausse question, plutôt demande de confirmation, ou forme d'hésitation, 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 mots les moins bien connus étaient prononcés légèrement ascendants ou peu descendants ou plats alors que les items connus étaient émis nettement descendants.Il est intéressant d'interpréter ces deux items, semi-interrogatif, puis énonciatif, 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 assertée ou éventuellement remise en question (CULIOLI).

La comparaison 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 interrogatives soit un décalage de la voix vers les zones aiguës. L'utilisation des divers niveaux de la tessiture à des fins linguistiques apparaît clairement ici.

Toutes les questions sans mot-outil sont de type " interrogation totale" (*Yes-No questions*) qui appellent, 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 interrogatif, 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 étudiées, il est clair que le questionnement avec la seule mélodie a un rendement maximal dans la période des premiers mots. Le trait commun à toutes ces questions mélodiques, outre leur contour ascendant, est le niveau élevé dans lequel se situe la voix, avec un Fo-m toujours supérieur à 470 Hz, et une culmination des énoncés dans le haut du niveau 4 ou dans le niveau 5. Ainsi les interrogatives ont le Fo le plus élevé de toutes les classes d'énoncés.

TABLEAU COMPARATIF

	QUESTIONS EN PROTO-LANGAGE	QUESTIONS ARTICULEES
Forme du contour	ascendant	ascenfant
M.Fo initial	408Hz	459Hz
Min. Fo	230Hz	310 Hz
M.Fo final	499Hz	ascendant [illegible] Hz
Max Fo final	720Hz	835 Hz

(M. = Moyenne)

[1] DE BOYSSON BARDIES, B.(1980), communication personnelle d'un rapport d'A.T.P. du C.N.R.S.. non publié.

[2] FONAGY, I.(1984),"La genèse de l'énoncé articulé", *Neuropsychatrie de l'Enfance* 32/10-11,517-527.

[3] HERMANN, E.(1942), "*Probleme der Frage* ", Göttingen : Vandenhoeck & Ruprecht.

[4] KASSAI, I.(1979),"Melodic patterns in child language", Budapest : *Magyar Fonetikai Füzetek*,4,147-170.

[5]KASSAI, I.(1987), "Early questions. preliminary report", Budapest : *Magyar Fonetikai Füzetek*, 17, 102-115.

[6] KONOPCZYNSKI, G.(1986), "*Du Prélangage au Langage : Acquisition de la Structuration Prosodique*," Thèse d'Etat, Université de Strasbourg II, vol. III, sous presse chez Buske Verlag, Hamburg.

[7] ROSSI M. & al (1981)," *L'Intonation De l'Acoustique à la Sémantique*", Paris : Klincksieck.

CONTEMPORARY CZECH PRONOUNCIATION: 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 auditorily analyzed by experienced listeners. A data structure for storing results of the auditory analyses 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 pronounciation 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 explicitely 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 point 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-syllabic 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: 5-33%), 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 consonants, weakening of articulation is here the most common change.

The number of mispronounced vowels is considerably higher than that of consonants: in 75 % of the speakers 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 consonants, omission and confusion of sounds. Eight speakers in our sample had a speech defect; in two other speakers the nasality was excessive. Regarding the frequency of errors in individual speech sounds: more than 10 % of errors were found in consonants *f, l, č, m, v* (in *f* and *l* more than 15 %), more than 5 % also *š, h, ž, c.*

In all, approx. 32 (1988: 36) % of all vowels were deformed.

In short vowels the most frequent deviation is a too open pronunciation, then comes a reduced timbre and changes in quantity (both shortening and lengthening).

In long vowels an open pronunciation and vowel shortening is very common. The most frequent deviation in consonants is their incomplete (weakened) realisation; the speech defects are found in sibilants and in the -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 text: 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 (8) 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 occasional mispronunciations may point to a certain instability in pronunciation;
- strangely enough, apart from the clear "logopedic cases", the famous Czech *ř* (*Dvořák*) remains unchanged.

A small table at the end of our paper gives some general results, showing sums and percentages of errors for individual classes of speech sounds. Again, a concentration of deviations in the data for vowels in comparison with those for consonants is apparent here in somewhat more detail. A correlation of these percentages with the results of the previous part of the analyses is high and significant ($r = 0.93$).

Considerable differences can be seen between the relative stability of the plosives, 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 respect to the position of the speech sounds within the text; all segments were coded, however, with respect to their occurrence 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 following facts have to pointed out:

- differences in numbers of deviations between initial, medial and final syllables: not only final syllables show, as could be expected, a higher number of deviations, but also the sounds in initital positions;
- no great differences were found in results for stressed vs. unstressed syllables.

In conclusion, two facts perhaps deserve to be mentioned again: firstly, a detailed analysis of our material reveals a picture radically different from the situation with which speech therapists of teachers of foreign students are confronted; secondly, the most common and widely spread are those mistakes originating in careless pronunciation habits, leading then to reduced intelligibility.

Numbers and percentages of mispronunciations in

Speech sounds	**N**	**Err**	**%**
Total:	14 520	2 473	17.0
Vowels (*total*):	5 940	1 899	31.9
Short:	4 686	1 591	33.9
Long:	1 254	308	24.5
Consonants (*total*)	8 580	574	6.6
Plosives:	2 343	90	3.8
Fricatives:	2 442	172	7.0
Affricates:	330	33	12.7
Nasals:	1 551	74	4.7
Sonorants:	1 914	205	8.3

N = number of sounds in a class
Err = number of mispronounced sounds
% = percentage of deviations (Err/N)

REFERENCES

FANT, G., L. HARD and A. KRUCKENBERG (1987)"Individual variations in text reading. A data-base pilot study", *RUUL* 17, 194-114

JANOTA, P. and Z. PALKOVÁ (1989) "Auditory analysis of speech segments" in *Proceedings of the 28th acoustics conference on physiology, acoustics, psychoacoustics, acoustic of music and of speech,* Štrbské Pleso, 194-197

PALKOVÁ, Z. (1990) "Stručná pravidla české ortoepie [Basic rules of Czech orthoepy], Prague, Czechoslovac Radio

STRATEGIES 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 70014, S-10044 Stockholm, Sweden
+Names in alphabetic order

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 analysed. Considerable variation in phrase and clause boundary realizations was observed. Strategies including both boundary and coherence signalling have been identified.

1. INTRODUCTION

This contribution represents 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 modelling 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) signalling and boundary (demarcative) signalling - 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 signalling 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 minimal pairs, where the location of the sentence internal clause 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 sturgeon, Putte peeps.)

3a. När han överlämnade sej, och bonden hälsade kungen med ett leende, så blev det bara så. (When he surrendered, and the farmer greeted the king with a smile, it just happened that way.)

3b. När han överlämnade sej och bonden, hälsade 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-internal boundaries.

3. SPEECH ANALYSIS

In the present speech corpus, considerable variation in the acoustic-phonetic signalling of phrasing and phrase boundaries was observed. Here we will not aim at giving an exhaustive description of the production data, but rather point to a few possible strategies in the exploitation of acoustic-phonetic cues for prosodic phrasing that we find especially interesting.

3.1. Boundary by duration only

One possible strategy is to use only duration for clause/phrase boundary signalling. This appears in some of the shorter sentences of our test material where there is no marking of the boundary in terms of F0. 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 F0 and duration (usually in combination) for the signalling 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 F0 (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 thus serves as a connective signal and may by itself be sufficient for the disambiguation of sentences 2a and 2b. Usually, however, this coherence signalling is accompanied by explicit boundary signalling. A typical F0 correlate here is the terminal F0-fall to a bottom F0 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 signalling without the use of deaccentuation. Here the F0 rise on "stör" followed by the F0 fall on "piper" together form a hat pattern [4], which serves as a connective signal. In this sentence we do not observe any obvious F0-boundary cues in connection with "fiskar", i.e. no F0-fall to a bottom level, although there are apparent segment lengthenings.

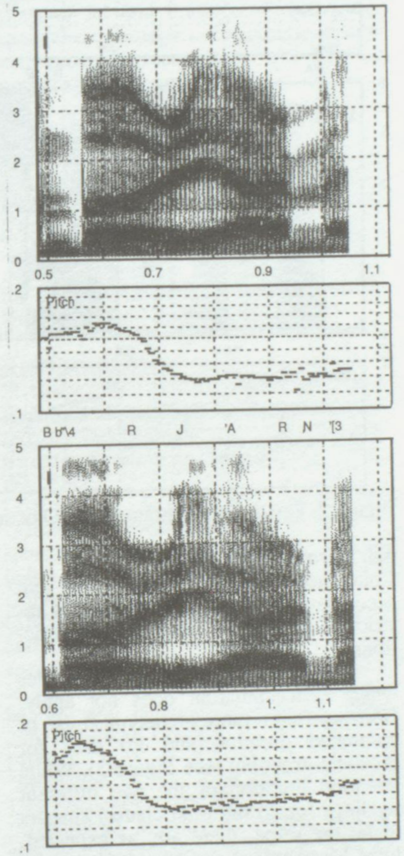


Figure 1. Partial spectrograms and FO of sentences 1a (top) and 1b (bottom)

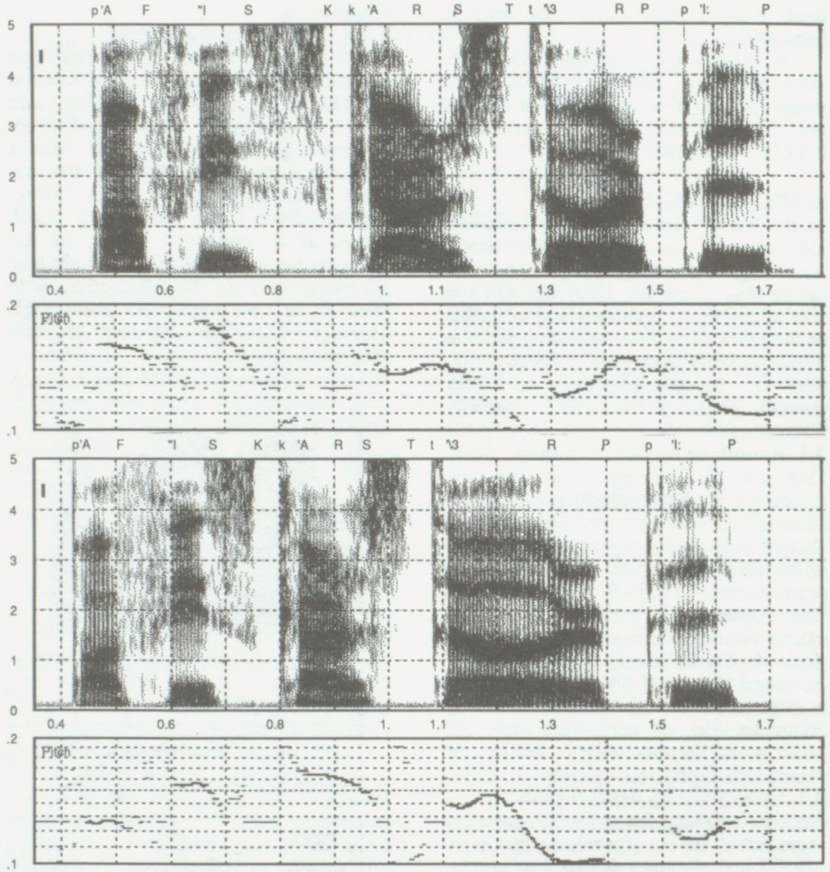


Figure 2. Partial spectrograms and FO of sentences 2a (top) and 2b (bottom)

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 combined 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,

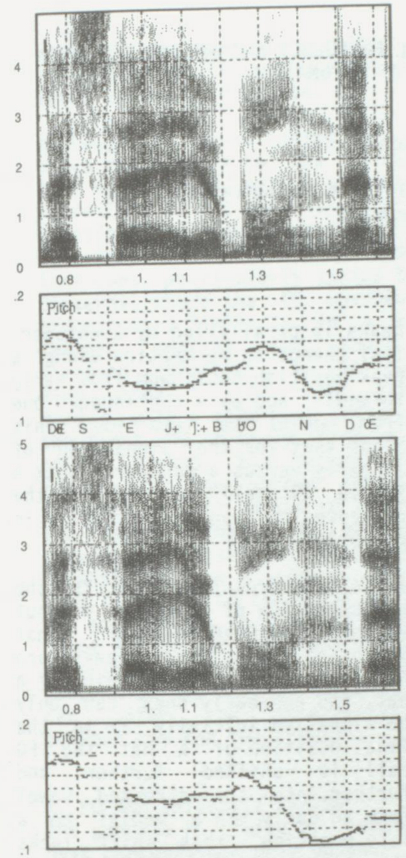


Figure 3. Partial spectrograms and FO of sentences 3a (top) and 3b (bottom)

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
[1]Bruce, G. (1985), "Structure and functions of prosody", in Guerin & Carré (eds.) Proceedings of the French Swedish Seminar on Speech, 549 - 559, Grenoble.

[2]Bruce, G. & Granström, B. (1990), "Modelling Swedish prosody in text-to-speech: phrasing", in Wiik & Raimo (eds.), Nordic Prosody V, 26-35, Turku University.

[3]Carlson, R., Granström, B. & Hunnicutt, S. (1990), "Multilingual text-to-speech development and applications", in Ainsworth (ed.), Advances in speech, hearing and language processing, 269-296, JAI Press, London.

[4]Hart, J. 't & Cohen, A. (1973), "Intonation by rule: a perceptual quest", J. of Phonetics 1, 309-327.

[5]House, D. & Bruce, G. (1990), "Word and focal accents in Swedish", in Wiik & Raimo (eds.), Nordic Prosody V, 156-173, Turku University.

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) 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 its second half or even the beginning of a subsequent unstressed syllable (late peak), which 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 parallelled 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 final 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 left 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 fall and rise timings, respectively, and the concomitant perception is equally continuous. Because of this, the early peak is the most natural F0 pattern at the end of a hat. It also accentuates the contrast between the low F0 in the stressed vowel and the high F0 level preceding it, thus adding to stress perception, which is weakened if the F0 fall is postponed and thus the high F0 level extended (figs. 1a, b).

Although the positioning of F0 peaks contributes to the perception of stressed syllables, this F0 feature is not the only factor. Durations of vowels and postvocalic consonants are also important cues, particularly inside hat patterns, where the F0 movements are minimal. Similarly, in a hat pattern uniting two abutting stressed syllables, as in 'Der Ring glänzt.' (*The ring glitters.*), with a late peak rise on the first and an early peak fall on the second, the segment durations in the second stressed syllable as well as the F0 timings are important for it to be perceived as stressed and thus differentiated from a single stress with late peak on the first syllable only (figs. 1b, c). In these cases we may ask to what extent intensity contributes to stress perception and whether changing it can alter the interpretation between one and two stresses.

2. F0 AND INTENSITY TIMING

The precise F0 timing of terminal peak contours not only depends on the peak type but also on the segmental structure of the stressed syllable. In medial peaks, the left-hand base point occurs at the beginning of the first consonant preceding the stressed vowel, the peak point at a time after vowel onset determined by 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 positioned where medial peaks have their left-hand base point; the right-hand base point occurs at the end of a lax (short) or about the centre of a tense (long) stressed vowel. 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 appr. 100 ms. To accommodate these F0 time courses in late peaks the stressed vowels are lengthened after the left base point, more so for lax 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, which weakens unstressed vowels and sonorants considerably, often reducing them to creaky voice and to irregular breathy glottal pulses. In late 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 'Sie hat ja gelogen.' (*She 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 diverge 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 plosive + syllabic nasal" as in 'Er ist ja geritten.' [... 'ɪtn̩] (*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 auditorily monitored. If a natural medial peak speech signal with its low final intensity in the above utterance is taken for LPC resynthesis 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 foregoing 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 in synthesis the perception is affected at the levels of voice quality and/or intonation. For pitch accents to be signalled effectively to 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. When 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

[1] CARLSON, R., GRANSTRÖM, B. & HUNNICUTT, S. (1990), "Multi-language text-to-speech development and applications", in *"Advances in speech, hearing, and language processing"*, Vol. 1 (W.A. AINSWORTH, ed.), London: JAI Press), 269-296.
[2] FRY, D. B. (1958), "Experiments in the perception of stress", *Language and Speech*, *1*, 126-152.
[3] KOHLER, K. J. (1987a), "Categorical pitch perception", in *"Proceedings of the XIth internainternational congress of phonetic sciences"*, Vol. 3, pp. 149-152, Tallinn: Academy of Sciences of the Estonian SSR.
[4] KOHLER K. J. (1987b), "The linguistic functions of F0 peaks", in *"Proceedings of the XIth international congress of phonetic sciences"*, Vol. 3, pp. 149-152, Tallinn: Academy of Sciences of the Estonian SSR.
[5] KOHLER, K. J. (1991), "Prosody in speech synthesis", *Journal of Phonetics*, *19*.

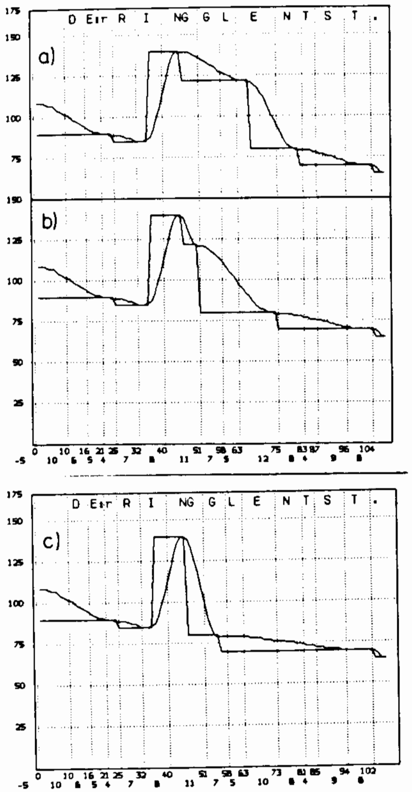


Fig. 1: Phonetic transcription and F0 (squares and cosine interpolations), in the German sentence 'Der Ring glänzt.' (RULSYS TTS); a) two stresses: hat pattern, late rise + medial fall, b) two stresses: hat pattern, late rise + early fall, c) one stress: late peak. Horizontal: cs frames (cumulative and for each segment), vertical: Hz.

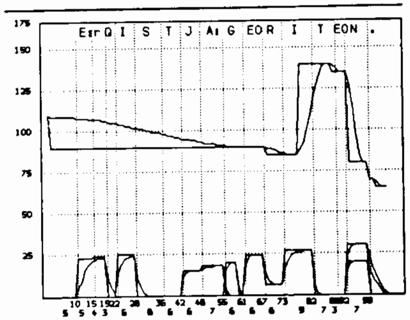


Fig. 2: Phonetic transcription, voice source A0 and nasal source AN (squares and cosine/2nd order interpolations) in the German sentence 'Er ist ja geritten'. with late peak (RULSYS TTS). Horizontal: cs frames (cumulative and for each segment), vertical: Hz for F0, dB for A0, AN.

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 (BP) 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'll do it I will" (cf. "If Tom will do it I 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 stacatto are perceptual impressions 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 present in 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 investigation 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 PROTOCOL

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 Klatt's [18] way of segmenting, combined with the works of Lehiste and Peterson [19, 25] which deal with the notions of onglides, offglides, 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 SPSS 4.1 for IBM VM/CMS 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 patterns of the sounds studied, regardless of the method. In 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 "pretonic", ST "stressed", and PST "posttonic". The values were kept in centimeters, but it suffices to multiply any value by 8. to obtain the corresponding value in milliseconds.

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 the word is in a prepausal position, confirming similar findings in what Klatt [18] called "prepausal lengthening". In the present study this syntactic or prepausal cue 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, in any context. BP in this study confirms again results from Simões [27] done with the extreme vowels [i,a,u] 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. Prepausal 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.

Table 1: ANOVA results of cell means and standard deviations by language, position and method.

MSp (C)V	Mean	Std Dev	BP (C)V	Mean	Std Dev
PR4	13.18	3.	PR4	12.	5.2
PR3	13.39	3.35	PR3	13.64	5.95
PR2	14.79	4.9	PR2	14.35	4.43
PR1	14.96	3.63	PR1	16.73	4.14
ST	17.97	5.63	ST	24.95	6.53
PST1	17.63	4.19	PST1	17.94	4.87
PST2	24.	1.92	PST2	16.38	4.92

MSp V	Mean	Std Dev	BP V	Mean	Std Dev
PR4	7.75	1.2	PR4	12.	6.36
PR3	7.61	2.61	PR3	7.38	1.16
PR2	8.22	3.10	PR2	8.58	2.83
PR1	8.09	2.19	PR1	9.66	2.91
ST	10.49	3.22	ST	15.78	5.59
PST1	9.8	3.15	PST1	10.29	3.96
PST2	14.14	2.01	PST2	9.75	4.29

Figure 1: ANOVA results of multiple range test. The symbol ◆ denotes pairs that are significantly different at the .05 level. Method-1 is indicated by (C)V and method-2, by V.

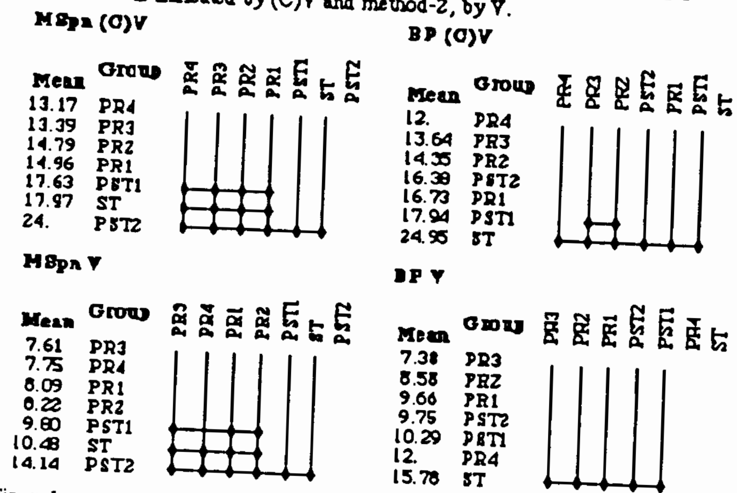


Figure 1 suggests a much greater regularity in temporal patterns in MSp than in BP. The pairing of groups (syllables) as seen in Figure 1 indicate quite a different behavior in MSp. In BP the stressed syllable seems to be a reference for the other syllables. In MSp the end of the word position, namely stressed and posttonic syllables seem to have a function of reference similar to the stressed position in BP. In other words, strong positions in MSp are more evenly distributed among syllables, especially the stressed and posttonic ones. This may be a result of syllables in general having relatively similar duration. Of course, the fact that pretonic, stressed and posttonic are different groups is still maintained in both languages as these results show. Figure 1 suggests that besides the inter-major group differences, there are intra-major group differences as well. Since these are statistical differences, definite conclusion will need perceptual analysis for validity and correct interpretation.

The notion of P-centers [7, 12, 16] has given the present analysis a clearer view of the temporal patterns observed. Although a perceptual analysis is necessary in the continuation of this study, the present results in Figure 1 from measurements at the acoustical level suggest that an increase in the number of measurements will provide a greater regularity in the temporal patterns of MSp. In the case of BP, the present results confirms the possibility of finding both types of rhythm. The possibility of finding both stress- and syllable-time rhythm in BP is not new. Abaurre-Gnerre [1, 2] explained this phenomenon in terms of "style", and Major [21, 22] in terms of a possible rhythmic change BP is undergoing presently. Major, however, concludes that BP is a stress-time language. Abaurre-Gnerre [2] suggests a more attractive explanation in terms of a rhythmico-stylistic criterion. Abaurre-Gnerre's solution is based on a scale that includes variation in language rhythm as one goes from a formal style (slow rate of speech) to a colloquial style (fast rate of speech). Paralleling this scale on style, the rhythm varies from syllable-time (formal) to stress-time (informal). Spanish and English are examples of languages on the extremes of the scale, i.e. syllable-time and stress-time respectively. It should be noted that in this scale, Peninsular Portuguese is placed between BP and English, namely with more stress-time characteristics, but still less than English. Another interesting aspect of Abaurre-Gnerre study is her attempt to link phonological processes to a type of rhythm. Vowel harmony, for example, may be related to a syllable-time rhythm. This can be extremely useful if such a relation can be established. If vowel harmony characterizes a syllable-time rhythm, this should not be a surprise because very often it indicates a more evenly distributed number of strong positions in a word. In other words, open vowels in BP only appear in strong position, i.e. stressed position. Vowel harmony in BP very often involves open vowels indicating a strengthening of the position where a closed vowel is realized as open.

In terms of a general theory of phonetics, the present study claims that rhythmic patterns may coexist in a given language and it is not limited to stylistic variation. Other factors may be present. Dialectical variation, for instance, may explain why one of the informants in Major [22], from Minas Gerais, may have the so-called syllable-time rhythm, or in terms of the present analysis, "vowel stability". This is my interpretation of the results in that work, which in a closer analysis suggest syllable-time characteristics instead of the proposed stress-time characteristic. The explanation presented here for these rhythmic alternations within the same language and intra-speaker, is that the speaker also manipulates rhythm at his/her will. The reasons are of a pragmatic nature where sometimes in the speaker-hearer interaction the speaker may feel a need for a clearer message.

4. REFERENCES

[1] ABAURRE-GNERRE, M.B. (1979), *Phonostylistic aspects of a Brazilian Portuguese dialect: implications for syllable structure constraints*, unpub. diss., Buffalo: State University of New York.

[2] ABAURRE-GNERRE, M.B. (1981), "Processos fonológicos segmentais como índice de padrões prosódicos diversos nos estilos formal e casual do português do Brasil", *Cadernos de estudos lingüísticos*, 2, 23-44.

[3] ABERCROMBIE, D. (1967), *Elements of general phonetics*, Edinburgh: Edinburgh University Press.

[4] BAUER, R.M. (1983), "Stress-timing and syllable-timing reanalyzed", *Journal of phonetics*, 11, 51-52.

[5] CÂMARA, J.M. (1970), *Estrutura da língua portuguesa*, Petrópolis, Brazil: Editora Vozes.

[6] CÂMARA, J.M. (1977), *Para o estudo da fonêmica portuguesa*, Rio de Janeiro: Padrão.

[7] COOPER, A.M., D.H. WHALEN & C.A. FOWLER (1986), "P-centers are unaffected by phonetic categorization", *Perception and psychophysics*, 39, 187-196.

[8] DELATTRE, P. (1963), "Research techniques for phonetic comparison of languages", *International review of applied linguistics in language teaching*, I, 85-97.

[9] FOWLER, C.A. (1979), "Perceptual centers" in speech production and perception, *Perception and psychophysics*, 25, 375-88.

[10] FOWLER, C.A. (1986), An event approach to the study of speech perception from a direct-realist perspective, *Journal of phonetics*, 14, 3-28.

[11] FOWLER, C.A., & L. TASSINARY (1981), "Natural measurement criteria for speech: the anisochrony illusion, *Attention and performance IX*, J. Long and A. Baddeley, eds., vol. 9, 521-35, Hillsdale, NJ: Erlbaum.

[12] FOWLER, C.A., D.H. WHALEN, & A.M. COOPER (1988), "Perceived timing is produced timing: a reply to Howell", *Perception and psychophysics*, 43, 94-98.

[13] GODÍNEZ, M. (1978) "A Survey of Spanish and Portuguese Phonetics", *UCLA Working papers in phonetics*.

[14] GREEN J.N. (1988), "Spanish", Martin Harris and Nigel Vincent (eds.), *The Romance languages*, New York: Oxford University Press, 79-130.

[15] HOWELL, P. (1984), "An acoustic determinant of perceived and produced anisochrony", Proceedings of the *10th International congress of phonetic sciences*, M.P.R. Van den Broecke & A. Cohen, eds., 429-33, Dordrecht, Holland: Foris.

[16] HOWELL, P. (1987), "Prediction of P-center location from the distribution of energy in the amplitude envelope: I", *Perception and psychophysics*, 43, 90-93.

[17] KELM, O.R. (1989) *Temporal aspects of speech rhythm which distinguish Mexican Spanish and Brazilian Portuguese*, unpub. diss., Berkeley, Ca.: University of California.

[18] KLATT, D.H. (1976), "Segmental Duration in English", *Journal of the acoustical society of America*, 59, 1208-21.

[19] LEHISTE, I. & G.E. PETERSON (1961), "Transitions, glides, and dipthongs", *Journal of the acoustical society of America*: 33, 268-77.

[20] LIEBERMAN, Ph. (1977), *Speech physiology and acoustic phonetics: an introduction*, New York: MacMillan Pub.

[21] MAJOR, R. (1981), Stress-timing in Brazilian Portuguese, *Journal of phonetics*, 9, 343-51.

[22] MAJOR, R. (1985), Stress and rhythm in Brazilian Portuguese, *Language*, 61, 2, 259-89.

[23] PARKER, E.M. and R.L. DIEHL (1984), "Identifying vowels in CVC syllables: effects of inserting silent and noise", *Perception and psychophysics*, 36(4), 369-80.

[24] PARKINSON, S. (1988), "Portuguese" Martin Harris and Nigel Vincent (eds.), *The Romance languages*, 130-69, New York: Oxford University Press.

[25] PETERSON, G.E. & I. LEHISTE (1960), "Duration of syllable nuclei in English." *Journal of the Acoustical Society of America*, 32, 693-703.

[26] PIKE, K.L. (1945), *The intonation of American English*, Ann Arbor: University of Michigan Press.

[27] SIMÕES, A.R.M. (1987), *Temporal organization of Brazilian Portuguese vowels in continuous speech: an acoustical study*, unpub. diss., Austin, TX.

[28] SIMÕES, A.R.M. (1987a), "Brazilian Portuguese rhythm: stress-time, syllable-time or samba?", paper presented at the University of Texas colloquium on hispanic and Luso-Brazilian Literatures, and Romance Linguistics, October.

[29] SIMÕES, A.R.M. (1990), "La enseñanza de los ritmos acentual y silábico", *Estudios de lingüística aplicada* - CELE/UNAM, año 8, 11, 130-47.

SYNTAX AND INTONATION IN ITALIAN NOUN PHRASES

S. Kori and H. Yasuda

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 a 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 tonal 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 noun phrases, F0 contours of six types of noun phases consisting of three content words 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 systematically varied (Table 1).

In all phrases, the number of inter-stressed syllables is kept constant at three. Three speakers from Northern Italy (SG 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 randomized 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 F0 course 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 second content word shows a more conspicuous protrusive movement than in sentence 1: the left-most content word in the right-branched constituent has a conspicuous tonal protrusion.

In speaker SG, however, the difference in branching construction is realized 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 (N & [N Adj]), 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 words than do 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 branched 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 EF, but not in all speakers. Moreover, in some sentences with left-branching constructions, there is a conspicuous F0 protrusion in the second content word. In fact, the difference in branching construction between sentences 6 (N & [N Adj]) 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) and sentence 5 ([N & N] Adj) 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 word. The inter-subject inconsistency found in sentence 4 (Adj [N & N]) could be interpreted as an interference between the mapping rule of 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 noun phrases consisting of three content words suggest 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 the 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

Table 1. Test noun phrases.
Content words are underlined and stressed syllables are italicized.

1.	[N Adj] & N	la *do*nna brasi*lia*na e il *bi*mbo
2.	N & [Adj N]	il ru*me*no e il *bra*vo brasi*lia*no
3.	[Adj N] & N	la *bel*la brasi*lia*na e il *bim*bo
4.	Adj [N & N]	i *gio*vani al*lie*vi e al*lie*ve
5.	[N & N] Adj	la *do*nna e il *bim*bo brasi*lia*ni
6.	N & [N Adj]	la *do*nna e il *bim*bo brasi*lia*no

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.

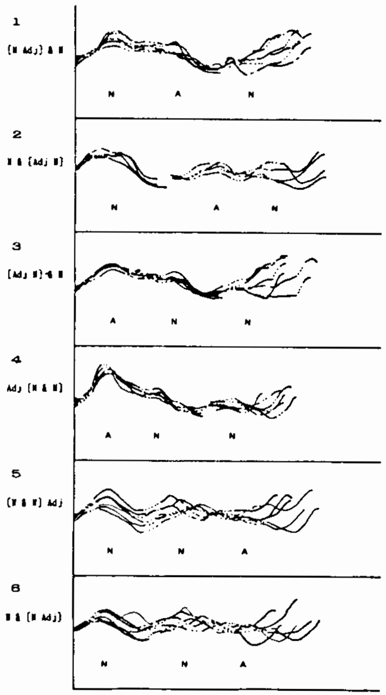


Figure 1.
F0 contours of test noun phrases
Speaker SG

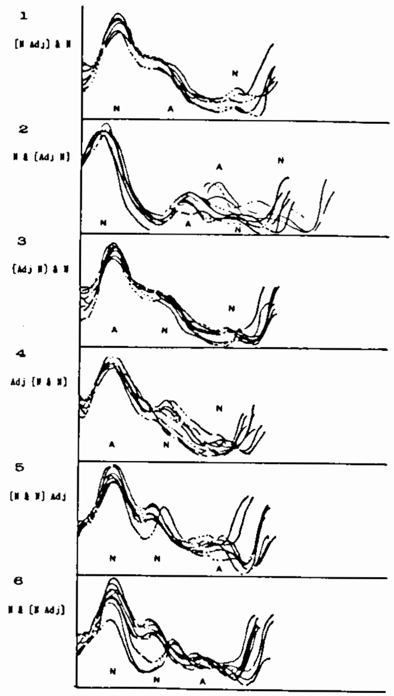


Figure 2.
F0 contours of test noun phrases
Speaker EF

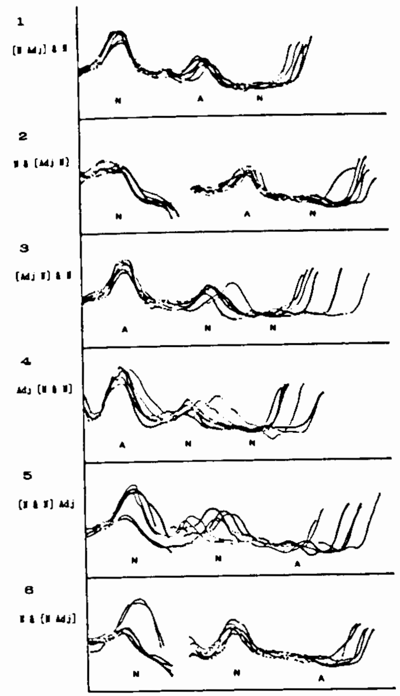


Figure 3.
F0 contours of test noun phrases
Speaker LT

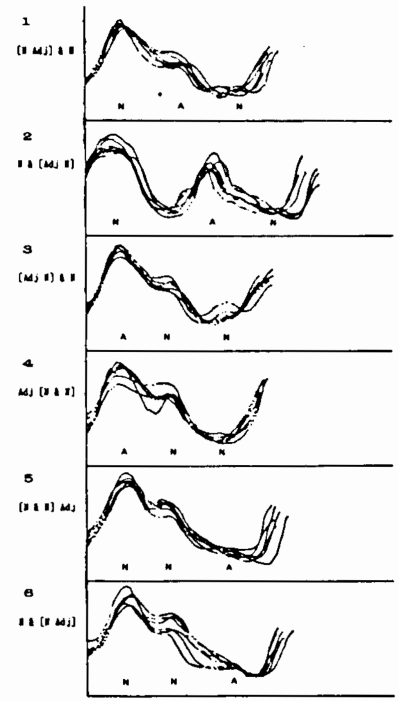


Figure 4.
F0 contours of test noun phrases
Speaker PC

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 value(s) 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 modalisation.

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 preterite. 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 five basic sets are :
I *The actor - the part*
II *The spider - the web*
III *The author - the book*
IV *The priest - the sermon*
V *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 or eacn 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 typed examples of the 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 two separate 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			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 1 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).

		1	2	3	4	5
I.	g - h	4	49	7	38	120
	h - i	24	41	40	118	112
II:	g - h	12	7	23	19	41
	h - i	25	27	35	103	103
III.	g - h	21	10	7	9	48
	h - i	15	16	35	29	128
IV.	g - h		6		25	67
	h - i	28	32	35	122	108
V.	g - h	5			65	26
	h - i	11	31	39	80	74
	i - i'				5	4

Fig. 1 shows the mean Fo differences on the verb in set II.

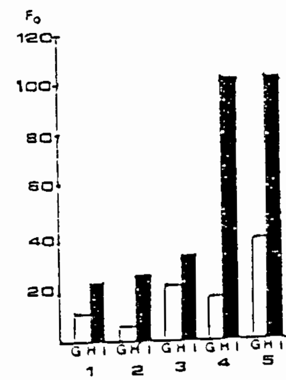


Fig. 1 Mean Fo differences on segment G-H and H-I for the 5 verbs in set II *The spider - the web*. Verbs: 1 *spun*, 2 *made*, 3 *wove*, 4 *spoilt*, 5 *tore*.

Fig. 2 shows the mean Fo differences on the verbs in each utterance of the 5 sets for all the informers.

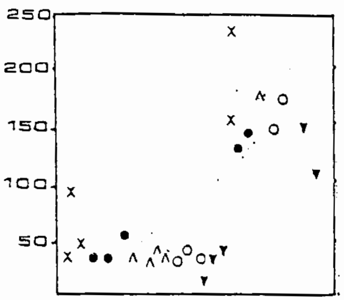


Fig. 2 Mean total Fo for segments G-H, H-I for the 5 utterances of the 5 sets : I x, II •, III ʌ, IV o, V ▾

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.

Fig. 3 gives two intonation contours illustrating a strong semantic association (in a) and a weak semantic association (in b) from set II produced by the same speaker.

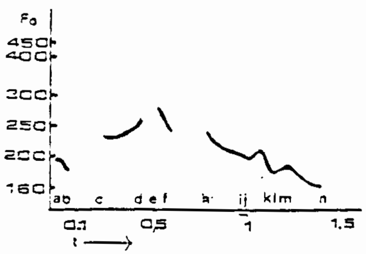


Fig. 3a :
A) The spider spun the web
AB CDEF GHI JK LMN

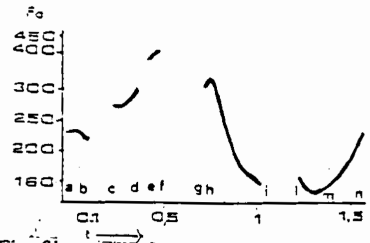


Fig. 3b :
The spider spoilt the web
AB CDEF GHI JK LMN

DISCUSSION and CONCLUSION

Both the perception tests and the instrumental analysis show that the differences perceived and produced on the five verbs representing varying degrees of semantic associations in each of the five sets of sentences were not gradual.

The verbs can be divided into two groups : group A, the verbs (1-3 in sets I, II, IV, V) that do not correspond to the prominent point of the utterance and correspond to small differences in Fo, and group B, the verbs (4-5 in sets I, II, IV, V) that correspond to the prominent point of the utterance and large differences in Fo. In set III the production of utterance 4 by all the speakers was similar to examples 1-3 in sets I, II, IV, V.

These results are in accordance with the polarity principle as analysed in the works of Edward Sapir, Roman Jakobson, Morris Halle and Harlan L. Lane.

The results can be explained within the framework of A. Culioli's linguistic theory of enunciative operations. The utterances in the corpus studied can be divided into two groups (A or B) depending on the variations in intonation on the verb in each utterance. In set II, *The spider - the web*, group A corresponds to the utterances with the following verbs : *spun, made, wove* and group B to those with the verbs : *spoilt, tore*. For group A the following definitions are possible : "A spider is a spinner, a maker, a weaver of webs". Whereas for group B the definition "A spider is a spoiler, a tearer of webs" is not possible within the framework of common acceptability.

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 utterances 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 modalisation.

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 P' is the modality of assertion (affirmative assertion for P, negative assertion for P'). The inability to choose between P and P' corresponds to the modality of interrogation.

The final fall corresponds to the choice of P or P' (assertion). The final rise corresponds to the point in the operation of modalisation at which the choice between P and P' 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

CULIOLI, A. (1991), *"Pour une linguistique de l'enonciation. Opérations et représentations"*, Paris. Ophrys.

JAKOBSON, R. and HALLE, M. (1956), *"Fundamentals of Language"*, The Hague : Mouton.

LANE, H. L. *"A Behavioral Basis for the Polarity Principe in Linguistics"*, Language. Vol. 43, N° 2, 494-511.

PETRYANKINA, V.I. (1987), *"Types of semantic relations between intonation and lexico-grammatical means (LGM) of language"*, Proceedings, XIth International Congress of Phonetic Sciences, Vol. 4, 267-270.

RIVIERE, C. (1983), *"Modal adjectives: transformations, synonymy, and complementation"*, Lingua 59, North-Holland Publishing Company.

SVETOZAROVA, N.D. (1987), *"Linguistic factors in sentence stress"*, Proceedings, XIth International Congress of Phonetic Sciences, Vol. 6, 110-113.

PERCEPTION OF INTONATIONAL 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 salience of intonation boundary after the wh-word, rather than the prominence of the focused wh-word per se.

1. 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 differs significantly in their prosodic domains as well. As a matter of fact,the difference 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), nani (what) etc. Incidentally, there are a class of words which are not wh-words but morphologically very similar to them: dareka (someone), dokoka (somewhere), nanika (something)etc. Those words are semantically marked, given their indefinite-pronounlike 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]$_{NP}$[mi-e'-ru]$_{VP}$
what-Nom. see-Pot.-Pres.
= What can (you) see?

(2)[na'nika]$_{NP}$ [mi-e'-ru]$_{VP}$
something see-Pot.-Prs.
= 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 expressed 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 Pierrehumbert & Beckman [1], the difference can be represented in terms of the difference of the 'intermediate phrase' defined as the domain of 'catathesis.' 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].)

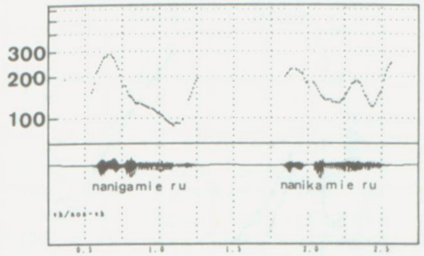


Fig.1 The F0 contours of wh question /naniga mieru / (left) and non-wh question /nanika mieru/ (right) as uttered by a male Tokyo Japanese speaker. The frequency scale is logalithmic.

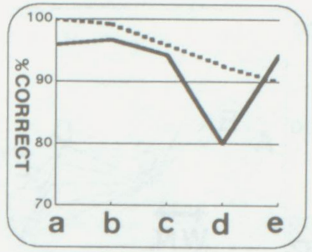


Fig.2 % correct identifications of wh question (real line) and non-wh question (dotted line). The abscissa represents the masking types indicated in the text.

(3) [na'niga mie'ru]
(4) [na'nika] [mie'ru]

In Fig.1, the peak F0 value of naniga is clearly higher than its nanika counterpart, and testifies to the presence of focus in the wh-word. This kind of focus-driven prominence in the wh-word is realized consistently, but it is by no means the only characteristic of wh-intonation. Rather, what makes the intonation shape of(1) visually distinct from that of (2) is the lack of salience of the prosodic boundary between NP and VP(a quick rise at the beginning of mieru). In short, there are two possible phonetic cues to the difference between (1) and (2): prominent F0 peak of the wh-word (Pw) and the salience of the prosodic boundary(Sb).

3. EXPERIMENT 1

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) nanigamieru
(1b) nanigamieru
(1c) nanigamieru
(1d) nanigamieru
(1e) nanigamieru
(2a) nanikamieru
(2b) nanikamieru
(2c) nanikamieru
(2d) nanikamieru
(2e) nanikamieru

In erasing 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

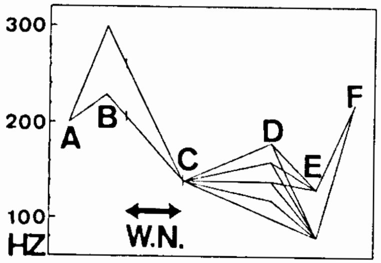


Fig.3 Schematic structure of the synthetic stimuli. Control points A-F were linearly interpolated as a gross approximation to natural intonations. The thick arrow indicates the time stretch masked with white noise.

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-question types. 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 full 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

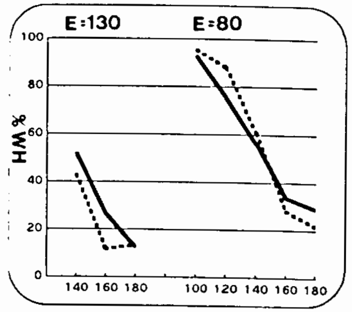


Fig.4 % wh-judgments of sixteen synthetic stimuli as the function of the D values (abscissa). Real lines stand for the stimuli with B=300Hz (prominent wh), and dotted lines stand for those with B=230Hz(not prominent).

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 230Hz. Point C stands for the beginning of the predicate mieru and is fixed at 140Hz. Point D is taken as representative of the cue Sb and is 180, 160, 140, 120 or 100Hz. Point E is the beginning of the sentence final rise and is either 130 or 80Hz. Point F is the target of the rise and is fixed at 220Hz. Of 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 PARCOR method, using the PANASYNS program developed by Hiroshi Imagawa and Shigeru Kiritani. 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 wh-question.The abscissa of the figure is a composite representation of D values for the stimuli with E=130 Hz (the leftward three values) and for the stimuli with E=80Hz (the rest). The real and dotted lines stand respectively for the stimuli 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-question,this effect is observed only when D is relatively high (180Hz or 160 Hz). Once D is set to relatively low values (120Hz or 100Hz), the stimuli were perceived mostly as wh-question irrespective of the B values.

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, it is noteworthy 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 that the D 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. Finally, 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 is observed in any pair of sentences having the same difference of focus placement as the one observed between (3) and (4).

6. REFERENCES

[1]Pierrehumbert,J.&M.Beckman (1988),Japanese Tone Structure. The MIT Press.
[2]Maëkawa,K.(1990), 'Muakusentohoogenno intoneeshon,' in Onseigengo, 4, 87-110.
[3]Fujisaki,H & H.Kawai (1988), 'Realization of linguistic information in the voice fundamental frequency contour of the spoken Japanese,'Ann.Bull.RILP,22,183-191.
[4]Kori,S.(1989), 'Kyoochooto intoneeshon,' in Sugito. M.ed. Kooza nihongoto nihongokyooiku,Vol.2. Tokyo, Meijishoin.

ACKNOWLEDGMENT

I am very grateful to Osamu Mizutani of NLRI for his comments on an earlier draft of this paper.

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 't Hart, Collier and Cohen (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 pitch 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, *excursion* 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 rises 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 posttonic parts (see below); for falls it is the *only* distinctive feature between accents Fl−/Fnl− and Fl+/Fnl+.
The *slope* of a pitch movement, expressed in semitones per second (ST/s), is the rate of change of F_0: a gradual or steep slope. Though not an independent distinctive feature in Russian, the rate of change of F_0 in combination with timing and/or posttonic part (see below) can differentiate between types of pitch accent (Odé 1989: 95).
The *posttonic part* is the syllable(s) immediately following the prominent syllable. Some pitch accents differ from one another on the basis of the level reached in this part: low vs. high vs. middle for rises; high vs. low (non-low) for falls.
The *pretonic part* is the syllable(s) immediately preceding the pitch accented syllable. The movement in a pretonic part can make the movement in the tonic syllable more salient.

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 (pretonic 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.

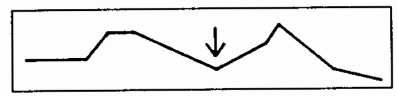


Figure 1: A turning point

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 F_0 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 F_0 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+, Fh−. Types Rh− and Rø− 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 naibolee* (Rl−) *často* (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+ has been observed in the utterance *nam nužno poechat' vot na jug* (Rl−) *s nej posidet'* (Fnl+) (230); and type Rl− followed by type Fnl− in *ty èto vy v polšestogo vstali i biletov* (Rl−) *ne chvatilo* (Fnl−) (252). Type Rl− is followed by Fl+ in *nu po-moemu* (Rl−) *raketa* (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 Rm−/+, 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 Fnl– did not occur after type Rh– in the corpus. In an experiment (Odé 1989:61-64) 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 repetiroval* (...) *scenu* (Rh–) *Sadka* (Rm–) *i ego ženy Ljubavy* (Fl+) (213); *no podvižki* (Rh–) *poka* (rm–) *mikroskopičeskie* (Fl–) (263).
In contrast to Nikolaeva's findings (1977:84), in my material there is no phonetic difference between types Rl–/Rh– in a final clause of a sentence and in non-final clauses. Both types occur in both positions, with different sizes of excursion, but the excursion is always large.
Type Rø–, which is frequently followed by a final fall, is in one case followed by type rl+ in the exclamation *čert* (Rø–) *ego znaet* (rl+) (282). An example of Rø– followed by type Rm+ is: *gm oni* (Rø–) *sotrudniki* (Rm+) *Akademii nauk* (231).
The final falls Fl–, Fnl–, Fl+ and Fnl+ have been found after one another, for example in afterthoughts: *v pjat'* (Rm–) *pjat'desjat* (Fl–) *ottuda* (Fl–) (249); *nam biletov ne chvatilo na ėtu* (Fl+) *raketu* (Fl+) (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: *a voobšče* (Fnl+) *vot tak vot v real'noj* (f) *žizni* (f) (231); *nu ėto Kolja Grinčenko* (Fnl+) *skazal* (Fnl+) (284); *ot nolja do pjati gradusov* (Fnl–) *tepla* (Fl–) (231).
Types Fnl– and Fnl+ are followed by the rises Rl–, Rh– and Rm+ in a few cases. Probably because of the high speaking rate in the spontaneous fragments no boundary was perceived after the fall. Examples are: *tut-to skazalas'* (Fnl+) *perestrojka* (Rl–) (256); *v tri časa idet bližajšaja* (Fnl–) *raketa* (Rl–) (249); *prjamo skažem nenormal'noe* (Fnl+) *raspredelenie* (f) *temperatury* (Rh–) (262); *značit priperlis'* (Fnl+) *tuda* (Fnl+) *v sem' utra* (Rm+) (251).
Finally, type Fh– can be followed by the same type: *ona* (...) *točnee* (Fh–) *sootvetstvovala* (Fh–) (219) and by type Fnl+: *da ėto* (Fh–) *dlja menja v obščem očen' suščestvennyj* (Fnl+) *vopros* (Fnl+) (233).
Type F^n+ 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 summarized into one *meaning*. If that is the case, the contextual functions found are *interpretations* of that meaning. Realizations of one type of pitch accent are *perceptually equivalent*, but contextual functions differ.
For example, type Fh– is interpreted as a question in *čto* (f) *ėto nam daet* (?) (Fh–). In the utterances *a obratno* and *i vozmožno* type Fh– is interpreted as the punctuation mark ':'. In the utterance *vospityvajte* (Rm+) *svoju mamu* (Fh–) *v takom duche* (Fh–), the stream of thoughts is incomplete and evokes a reaction. On the other hand, in questions and in incomplete utterances we also find type Rl–, e.g. in *oni studenty* (?) and *oni uechali ottuda* (...).
At the congress more examples of combinations of pitch accent will be presented with their interpretation.

6 REFERENCES

't HART, J., COLLIER, R., COHEN, A. (1990), *A perceptual study of intonation: An experimental-phonetic approach to speech melody*, Cambridge.
NIKOLAEVA, T.M. (1977), *Frazovaja intonacija slavjanskich jazykov*, Moskva.
ODÉ, C. (1989), *Russian Intonation: A Perceptual Description*, Amsterdam.
ROOIJ, J.J.DE (1979), *Speech punctuation, an acoustic and perceptual study of some aspects of speech prosody in Dutch*, Doct. Diss., Utrecht.

Table 1. Types of rising pitch accent: average values and maximum and minimum values (limits of perceptual tolerance). R = rise with large excursion, r = rise with normal excursion, l = low posttonic part, h = high posttonic part, ø= no posttonic part, m = middle posttonic part, - = early timing, + = late timing.

type	excursion	timing	posttonics	slope	register	picture
Rl-	17 ST (13-21)	89% early 11% late	low	76 ST/s (54-116)	high	
Rh-	17 ST (15-20)	95% early 5% late	high	74 ST/s (35-120)	high	
Rø-	16 ST (13-21)	84% early 16% late	ø	73 ST/s (30-86)	high	
Rm-/+	15 ST (11-17)	70% early 30% late	middle	54 ST/s (39-94)	high	
rm-/+	10 ST (8.5-12)	60% early 40% late	middle	35 ST/s (19-56)	low	
rl-/+	11 ST (9-12)	87.5% early 12.5% late	low	52 ST/s (23-95)	low	

Table 2. Types of falling pitch accent: average values and maximum and minimum values (limits of perceptual tolerance). F = fall, l = low: the lowest level of the speaker is reached in the movement, nl = non-low: the lowest level is not reached, h = high posttonic part, f = fall to a level above non-low, ^{n} = the configuration is repeated, - = early timing, + = late timing.

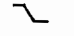
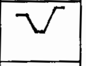
type	excursion	slope	above low	posttonics	picture
Fl-	8 ST (6-11)	47 ST/s (39-71)	0 ST	low	
Fnl-	7 ST (3-16)	42 ST/s (15-62)	4 ST (3-5)	non-low	
Fl+	9 ST (6-13)	47 ST/s (32-60)	0 ST	low	
Fnl+	8 ST (5-11)	50 ST/s (26-83)	4 ST (2-5)	non-low	
Fh-	6 ST (5-7)	35 ST/s (10-58)	4 ST	rises 9 ST to 13 ST above low (7.5-17)	
F^n+	10 ST (8-13)	65 ST/s (55-73)		rising	
f-/+	4 ST (2-6)	28 ST/s (14-57)	6 ST (5-8)	varying	

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 accents Rm-/+, rm-/+, rl-/+ and f-/+ 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 indication '-/+' has been left out of this table. Single cases are indicated with the sign 0.

	Rl-	Rh-	Rø-	Rm	rm	rl	Fl-	Fnl-	Fl+	Fnl+	Fh-	F^n+	f
Rl-				0	0			0	0	0			
Rh-		0		0	0		x		x	x			
Rø-				0	0	0	x		x	x			0
Rm	x	x	x	x	x	x	x	x	x	x	x		x
rm	x	x	x	x	x	x	x	x	x	x	x		x
rl										0			
Fl-							0						
Fnl-	0						0						
Fl+									0				
Fnl+	0			0	0					0			0
Fh-										0	0		
F^n+													
f	x	x	x	x	x	x	x	x	x	x	x		x

A CROSSLINGUISTIC DESCRIPTION OF INTONATION CONTOURS OF A MULTILANGUAGE TEXT-TO-SPEECH SYSTEM

G. Olaszy

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 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, German, Italian, Esperanto and Finnish. New languages can be adapted easily to the basic system. Dutch and Spanish are under development. The synthesis hardware is the PCF8200 formant synthesizer. In MULTIVOX a modular representation of intonation has been implemented.

2. ELEMENTS FOR INTONATION AND STRESS

In devising acceptable intonation for unrestricted 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 correction 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)(ms)
8. End point (E) of the pitch contour
9. Lengthening (L) of the stressed vowel

The degree parameter 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 1.

Unit	Degree High	Medium	Low	
S/E	125	110	95	Hz
R/F	25	15	5	%
N	–	–	–	
St	2	0,5	0,25	Hz/ms
L	3x	2x	1,5x	times

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 cooccurs with lengthening or not? This and the place of the accent is shown in Table 2.
Table 2. expresses that in Italian and in 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.

Table 2.

Language	Pitch change	Lengthening	Accented syllable
Hungarian	+	–	initial
German	+	–	any
Italian	+	+	any
Esperanto	+	+	penultimate
Finnish	+	–	initial

(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)(StH)+(FM)(StM)EM
for a long vowel (VV):
SM(RM)(StM)+N(x)+(FM)(StM)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. Word stress categories:

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 5 types of rules serve for word stress realisation in the mentioned five languages.

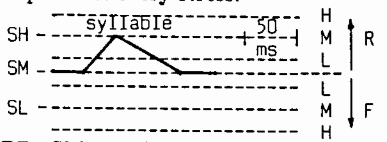
3.2. Algorithms for stress assignment

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 can be determined by the rules 1 and 3. If the stress is signalled by diacritics – 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] 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 determination.
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*, *Komponist*, *studieren*).
D3.An unstressed prefix is followed by a stressed syllable (*bekommen*, *gesagt*).
D4.In two syllable words the long vowel (if there is any) is stressed (*fahren*, *sehen*, *primär*), else the first (*Silbe*, *Tausend*). This last rule is based on empirical observations.
Using these rules for finding the place of stress in German words a correct pitch superimposing is performed in 95% of the cases. The evaluation of these rules were done by listening to 1600 German sentences [8] and 50 text files (one A4 page each) gathered from books and newspapers. A weaker point of the German word stress assignment is the case of compound words. 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.

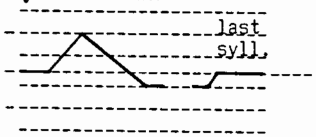
3.3. Pitch patterns for word stress

The following types of pitch patterns (PP) are used to create the frequency component of stress:

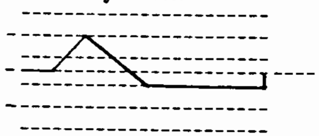
PP1.SM+RM(StM)+FH(StH)+EM
Hungarian: first syllable,
Italian: every stress except final,
German: stressed suffix
Esperanto: every stress.



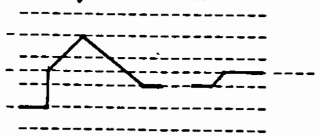
PP2.SM+RM(StM)+FH(StH)+N(x)+ +RL(StM)+EM
German: first syllable in more-than-two-syllable words.



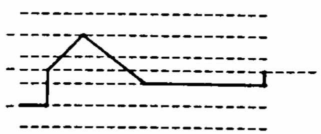
PP3.SM+RM(StM)+FH(StH)+N(x)+ +Ju(EM)
German: first syllable in two-syllable words,
Finnish: every stress.



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 closer to the natural variation among stressed, unstressed, and neutral parts 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 for every language in the system. In sum, concerning word stress generation three types of cases are used: stressed, unstressed and neutral sequences. All these patterns 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 in the system. 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 rised continuously in the last two syllables of the phrase e.g. RL(StM). The pitch is set back (JdM) during the phrase pause 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 listening to long texts should not be uncomfortable [2].

Two types of sentence intonations are generated automatically in the MULTIVOX system: one for declarative sentences and one for questions. For declaratives the general theoretical pattern is a linear falling one. This pattern is used for 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 3.

Sentence length		Start pitch	Steepness
very short	(300ms)	120 Hz	10Hz/100ms
short	(600ms)	118	3
medium	(1 s)	116	2
normal	(3 s)	114	0.5
long	(8 s)	112	0.2
very long up to	(15 s)	110	0.1

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 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 pattern is used for all the languages in the system. A high peak is set on the Q word i.e.

RH(StH)+ FH(StM)+ FL(StM)

and afterwards a 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 was. The place of pitch change depends on the Q word and 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 pattern is a peak i.e.

RH(StH)+FH(StM)

on the penultimate syllable.

5.3. One-syllable questions

The pattern is the same for all the languages for one-syllable questions. This is a rising one i.e.

SL+RL(StL)+RL(StM)+RH(StH).

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

[1] ALLEN, J.–HUNNICUTT, M.S.–KLATT, D. (1987), *"From text to speech. The MITalk system"*, Cambridge.

[2] COLLIER, R. (1990), "Multilingual intonation synthesis: principles and applications", *Proc. of the ESCA Workshop on Speech Synthesis*, Autrans, 273–76.

[3] KOHLER, J.K. (1990), "Improving the prosody in German text-to-speech output", *Proc. of the ESCA Workshop on Speech Synthesis*, Autrans, France, 83–87.

[4] OLASZY, G. (1989), "Speech synthesis in Hungary from the beginnings up to 1989", *Proc. of the Speech Research '89 Conference*, Budapest, 289–92.

[5] OLASZY, G. (1991), "Timing algorithms in the MULTIVOX text-to-speech system", In: *Temporal patterns of speech.* Ed. Gósy, M. Budapest.

[6] O'SHAUGHNESSY, D.D. (1989), "Parsing with a small dictionary for applications such as text to speech", *Computational Linguistics* Vol.15 num. 2., 97–108.

[7] SALAZA, P.L. (1990), "Phonetic transcription rules for text-to-speech synthesis of Italian", *Phonetica* 47, 66–83.

[8] SOTSCHEK, J. (1984), "Sätze für Sprachgütemessungen und ihre phonologische Anpassung an die deutsche Sprache", *Fortschritte der Akustik, DAGA'84*, 873–876.

[9] TERKEN, J.M.B.–COLLIER, R. (1989), "Automatic synthesis of natural-sounding intonation for text-to-speech conversion in Dutch", *Proc. of Eurospeech'89*, Edinburgh, 357–359.

MEASURING INTONATION AT LOW SIGNAL-TO-NOISE RATIOS

V.Pikturna

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 glob 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 shaping of the intonation contour is 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 femail speaker is 450 Hz then the frequency region 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 principle 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 its shape 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:

$$H(z)= =20\lg\left|1+a_1z^{-1}+a_2z^{-2}+a_3z^{-3}\right|^{-2}$$

$a(i)$ being the LP coefficients, $z=exp(-j\omega)$, ω being the current frequency. The value of shifting downwards 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 a 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 $F(k)$, the frequencies and $A(k)$, the levels of maxima of spectral peaks. Obviously, k is not always a number of a pitch harmonic. We have chosen a method of evaluating PF most close to 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 $A(k)$ 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 correct value 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.[1]). 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...15) 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. Flattness of the spectrum

The slope of spectra of the white noise computed from short frames is much less than that of voiced sounds [2]. The dynamic range Δ of the ATH shows to be the very efficient measure of the spectral flattness. We formulate the following feature: a frame is unvoiced if

Δ<10 dB when SNR>10 dB,
Δ< 7 dB when SNR<10 dB

(Fig.1, V/UV2).

4.3. Number and disposition of harmonics

If the processing of spectrum results in finding less than 3 spectral peaks, the frame is labeled unvoiced (Fig.1, V/UV3).

If examining of three peaks in the region of spectral energy maximum does not result in finding equidistancies, the frame is labeled unvoiced (Fig.1, V/UV4).

5. SHAPING OF THE INTONATION CURVE

5.1. Jumps to a neighbouring harmonic

To avoid jumps to the 2nd or to the 0.5th harmonic, the past of the intonation curve is used: the current value of the PF is compared to the average of all previous non-zero values of the PF. If it exceeds twice or is twice less than the average mentioned, it is devided (multiplied) by 2. If the declination is greater than 2 times, the PF is set to zero. We find such an approach more effective than one-step-back control.

5.2. 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 femail) using a limited speech material. When using knowledge of a human expert, the intonation curve remains at SNR down to 0 dB.

7. REFERENCES

[1]ALLIK,J.,MIHKLA,M., ROSS, J.(1984),"Comment on "Measurement of pitch in speech:An implementation of Goldsteins theory of pitch perception", *J.Acoust.Soc.Am.*,75(6),1855-1857.

[2]PIKTURNA,V., RUDŽIONIS,A. (1990),"Pitch measuring from spectra of noisy speech:amplitude thresholding versus identifying of harmonics", *Proc. 3rd Australian Int. Conf. on Speech Science and Technology*, Melbourn, 6 p.

[3]RABINER,L.R.,SAMBUR,M.R., SCHMIDT,C.E. (1975), "Applications of a nonlinear smoothing algorithm to speech processing", *IEEE Trans. Acoust.,Speech and Signal Processing*,23,554-557.

[4]SREENIVAS,T.V.,RAO,P.V.S. (1979),"Pitch extraction from corrupted harmonics of the power spectrum", *J.Acoust.Soc.Am.*,65,223-228.

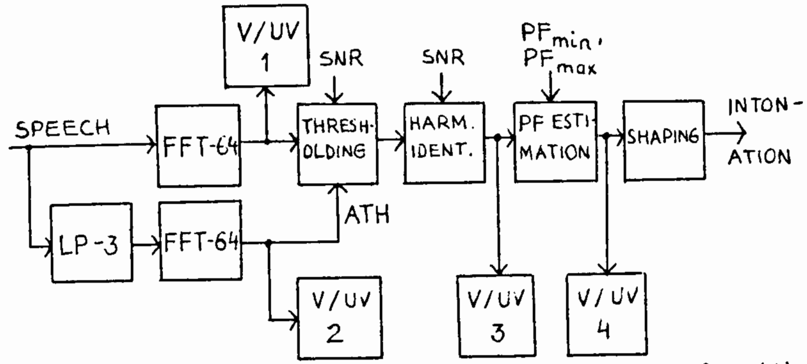


Fig.1. Block diagram of the intonation measuring algorithm

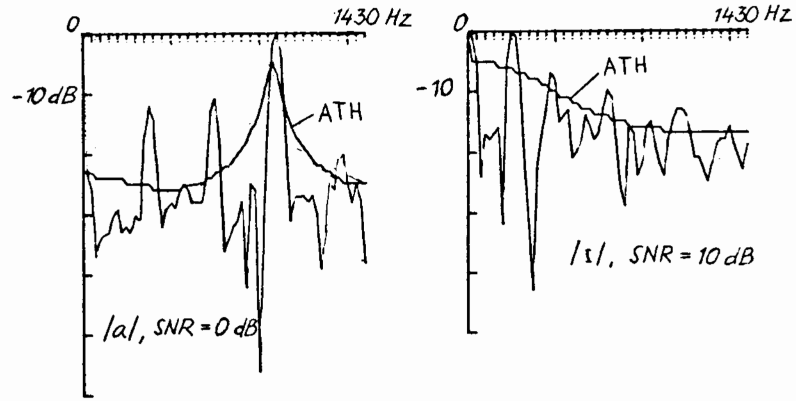


Fig.2. Effect of the adaptive threshold ATH

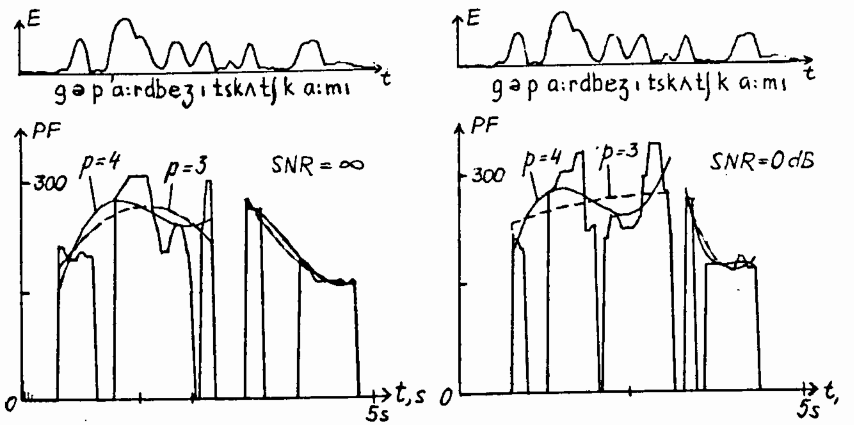


Fig.3. Intonation curves approximated by the 3rd and 4th order polynomials

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 an 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 speech 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 portions [5]. Aperiodicity of voiced speech can be due in some cases to severe irregularity in occurence of glottal pulses. In such cases it is impractical to define f0 in terms of *production* (as the inverse of the interval between glottal pulses), and it may seem preferable to define it instead in terms of *perception* (pitch).
Several perception-based methods have been proposed [6, 7, 8], most of which are based on the pitch perception theories of Goldstein or Terhardt [9, 10, 11]. The general principle shared by these models is that pitch is determined from a spectral pattern by searching for a common subharmonic of major spectral components. The spectral pattern is presumably produced by peripheral analysis in the cochlea, and the matching of subharmonics carried out at a more central stage. Spectral pattern matching theories are being questioned of late, because physiological data support alternative theories that assume that pitch derives instead from the *periodicity of neural discharges*.

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. Licklider's ideas have been developped recently by other authors [12, 13, 14, 15, 16]. Autocorrelation, as used in Licklider's model, does not *require* a filtering stage: it can be performed directly on the raw speech signal [4]. This raises a question: what might be the advantage of peripheral filtering for pitch perception? One can imagine several possible answers:
a) The signal-to-noise ratio or the periodicity might be better within a restricted group of channels. [17][18].
b) Small differences of phase from period to period can result in large differences in wave shape, causing a comparison method such as autocorrelation to fail. Filtering might reduce such interaction.

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 called pitch is not the same object as speech fundamental frequency (often also called pitch) and the tasks of extracting the former or perceiving the latter are not equivalent.

2. METHODS

2.1. Database

Data was taken from an f0 database developped at ATR [19, 20]. The speech was sampled 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 voiced frames at a 400 Hz frame rate. The f0 values cover a 2-octave range centered on about 125 Hz.

2.2. AMDF

All experiments are base on the Average Magnitude Difference Function (AMDF) method [21]. The AMDF is defined as:

$$\mathrm{AMDF(lag)} = \int_{\mathrm{window}} |S(t) - S(t + \mathrm{lag})|\, dt$$

The lag at the first major dip indicates the period. The AMDF produces as a by-product a parameter that can be interpreted as a *measure of periodicity*. This is defined as:

$$\mathrm{PM} = \log_2\left(\frac{\mathrm{mean(AMDF)}}{\mathrm{AMDF(period)}}\right)$$

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:

$$h(t) = A(t - T_l)^{v} \exp(-(t - T_l) / T_f) \; \sin(2\pi F(t - T_l))$$

where F is the characteristic frequency, T_l 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.

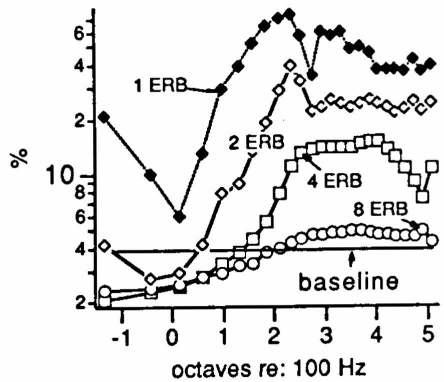


Fig. 1. Error rate as a function of center frequency for various channel bandwidths measured in ERBs.

3. EXPERIMENTS

3.1. Baseline

The error rate of "vanilla" AMDF over the database is 3.84%.

3.2. Individual revcor 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 revcor channel outputs 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.

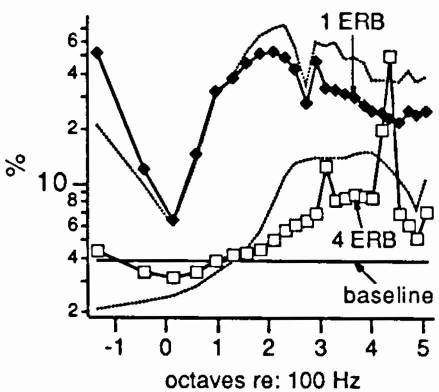


Fig. 2. Error rates for half-wave rectified revcor filter outputs. Dotted lines: rates for raw outputs.

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 revcor 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 interaction 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 1: error rates for various degrees of smoothing:

window size:	10 ms	20 ms	40 ms	80 ms
zero at:	1 kHz	500 Hz	250 Hz	125 Hz
error rate (%):	3.19	2.44	2.74	3.96

Given this simple result, one might be tempted to apply low-pass filtering systematically. This would be unwise 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 strategy appears to be to combine information accross channels. Simple addition of AMDF patterns yield 2.9 % errors for a 1 ERB bandwidth. This is in striking contrast with the rates 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). Uniform weights for all channels, as obtained by amplitude normalization, proved disappointing (5.15% for 1 ERB bandwidth).

CONCLUSION

An f0 extraction method based that splits the speech signal over a filter-bank before calculating the AMDF within each channel and combining the patterns improves reliablity 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 multiple-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 author wishes 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 revcor filter software.

BIBLIOGRAPHY

[1.] Licklider, J. C. R. (1956), "Auditory frequency analysis", *Information theory*, Cherry ed. Butterworth: London, 253-268.

[2.] Licklider, J. C. R. (1959), "Three auditory theories", *Psychology, a study of a science*, Koch ed. McGraw-Hill: 41-144.

[3.] Licklider, J. C. R. (1962), "Periodicity pitch and related auditory process models", International Audiology. *1*, 11-36.

[4.] Hess, W. (1983), *Pitch determination of speech signals*, Springer-Verlag: Berlin. Pages.

[5.] Hedelin, P. and D. Huber (1990), "Pitch period determination of aperiodic speech signals", *IEEE-ICASSP*, 361-364.

[6.] Duifhuis, H., L. F. Willems and R. J. Sluyter (1982), "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception", JASA. 1568-1580.

[7.] Hermes, D. J. (1988), "Measurement of pitch by subharmonic summation", JASA. *83*, 257-264.

[8.] Scheffers, M. T. M. (1983), "Sifting vowels",

[9.] Goldstein, J. L. (1973), "An optimum processor theory for the central formation of the pitch of complex tones", JASA. *54*, 1496-1516.

[10.] Terhardt, E. (1974), "Pitch, consonance and harmony", JASA. *55*, 1061-1069.

[11.] de Boer, E. (1977), "Pitch theories unified", *Psychophysics and physiology of hearing*, Evans and Wilson ed. Academic: London, 323-334.

[12.] Moore, B. C. J. (1982), *An introduction to the psychology of hearing*, Academic Press: London. Pages.

[13.] van Noorden, L. (1982), "Two channel pitch perception", *Music, mind, and brain*, Clynes ed. Plenum press: London, 251-269.

[14.] Lyon, R. (1984), "Computational models of neural auditory processing", *IEEE ICASSP*, 36.1.(1-4).

[15.] de Cheveigné, A. (1986), "A pitch perception model", *Proc. IEEE ICASSP*, 897-900.

[16.] Meddis, R. and M. Hewitt (1988), "A computational model of low pitch judgement", *Basic issues in hearing*, Duifuis, Horst and Witt ed. Academic: London, 148-153.

[17.] Fujimura, O. (1968), "An approximation to voice aperiodicity", IEEE Trans. Audio and Electroacoustics. *16*, 68-72.

[18.] Rodet, X., P. Depalle and G. Poirot (1988), "Diphone sound synthesis based on spectral envelopes and harmonic/noise excitation functions.", Proceedings of the ICMC, Köln (RFA), 313-321.

[19.] Kuwabara, H., Y. Sagisaka, K. Takeda and M. Abe (1989), "Construction of ATR Japanese speech database as a research tool", ATR technical report TR-I-0086.

[20.] Abe, M. and H. Kuwabara (1989), "Pitch frequency database on continuous speech", ATR technical report TR-I-0078.

[21.] Ross, M. J., H. L. Shaffer, A. Cohen, R. Freudberg and H. J. Manley (1974), "Average magnitude difference function pitch extractor", IEEE Trans. ASSP. *22*, 353-362.

[22.] Holdsworth, J., I. Nimmo-Smith, R. D. Patterson and P. Rice (1988), "Implementing the GammaTone filter bank",

[23.] Moore, B. C. J. and B. R. Glasberg (1983), "Suggested formulae for calculating auditory filter bandwidths and excitation patterns", JASA. *74*, 750-753.

[24.] Carney, H. and T. C. T. Yin (1988), "Temporal coding of resonances by low-frequency auditory nerve fibers: single fiber responses and a population model.", J. Neurophysiol. *60*, 1653-1677.

[25.] de Cheveigné, A. (1990), "Experiments in pitch extraction.", ATR Technical report TR-I-0103, 39p.

[26.] Johnson, D. H. (1980), "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones", JASA. *68*, 1115-1122.

[27.] Lyon, R. F. (1983-1989), "A computational model of binaural localization and separation", *Natural computation*, Richards ed. MIT Press: Cambridge, Mass, 319-327.

[28.] Meddis, R. and M. J. Hewitt (1990), "Modelling the identification of concurrent vowels with different fundamental frequencies", Submitted for publication.

A COMPUTER ASSISTED METHOD OF INVESTIGATING INTONATIONAL CORRELATIONS IN ADJACENT UTTERANCES

SAPPOK, Ch., KASATKINA, R., KODZASOV, S.

RUHR UNIVERSITÄT BOCHUM, BRD
INSTITUT RUSSKOGO JAZYKA AN SSSR, MOSKVA

0. Abstract
The intonation of adjacent turns in dialogs conveys information of at least two types: it indicates the sentence type of the utterence 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 represented 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 digital technology to make an entire communicative situation repeatable and subject to modification, in a way similar to the propositions formulated by HERTRICH and GARTENBERG 1989. The evidence we will adduce will prove to favour one of the following alternatives: Can we ascribe 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 vs. discourse 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 case of the absence of sufficient information, the modal utterences and their adjacent combinations are interpretable, since there appears to be a set of "default" knowledge.

2. The Method
The material consists of 12 microdialogs consisting of 4 turns each, and a preceding description of the situation. As to the organisation of the material in the form of a data base cf. SAPPOK 1990. The situation consists of a variable combination of propositions, the dialogs having always the same lexical form, as can be seen in the samples shown in Chart 1.

	S+R	R+S
poly vymyty		
1.1. A+,B+,A»B	a	1
1.2. A+,B-,A»B	b	2
1.3. A+,B±,A»B	c	3
1.4. A-,B-,-»?	d	4
3.1. A±,B±	i	9
3.2. A+,B-,?rad	j	0
poly vymyty		
	S+R	R+S
kleenka isporchena		
2.1. A-,B-,A»B	e	5
2.2. A-,B+,A»B	f	6
2.3. A-,B±,A»B	g	7
2.4. A+,B+,+»?	h	8
4.1. A±,B±	k	q
4.2. A-,B+,?rad	l	w
kleenka isporchena		
	S+R	R+S

Chart 1. Correspondences of attitudes, situations and symbols (as described in the text).

The description of the situation and the text of the dialogs were presented visually in written form to pairs of Russian native speakers who performed them orally 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 KNIPŠIL'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+R to R+S. The following symbols show keys to be pressed, after which the resulting dialog can be heard.

Ситуация 1.1.

А. и В. хорошо относятся к Ивановой.

А. хочет усилить это отношение.

А. – Ты замечаешь, что полы вымыты?

В. – Да-а. А кто это сделал?

А. – Иванова.

В. – Иванова?

А. – Да, Иванова.

Ситуация 1.2.

А. хорошо относится к Ивановой, а В. плохо. А. хочет изменить отношение В. к Ивановой на хорошее.

А. – Ты замечаешь, что полы вымыты?

В. – Да-а. А кто это сделал?

А. – Иванова.

В. – Иванова?

А. – Да, Иванова.

Chart 2. Instructions for two microdialogues as presented to the speakers

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 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

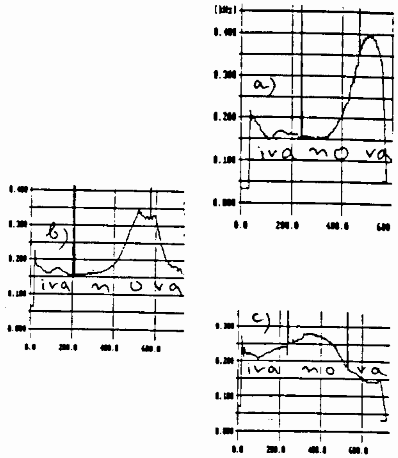


Chart 3. a) - c) Three reactions of a subject to neutral, positive and negative utterances in the dialogue with the computer.

three different questions of type A2 as reactions to B1 utterances of neutral, positive and negative versions. Although the subject has no explicit information about the nature of these turns beforehand, he reacts in a way comparable to the versions with explicit information.

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 conformity of expectations or their disparity as perceived by the respondent.
- 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 judgements 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 can 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 intervening 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 judgement "not similar" is slightly diminished in the case of 2.3. and 2.1., where the partner's attitude is neutral and negative, respectively. The intervening question A.1. is classified as intense ("a high degree of interrogativity"), 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 availible. Preliminary interpretation shows that
- the reaction of the subjects to the situations and dialogs 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:
[1] HERTRICH, I., GARTENBERG, R.D. (1989): A new method in intonation research using partly controlled, simulated dialogues. in J.P.Tubach, J.J.Manani (eds.): Eurospeech 1989, European Conference on Speech Technology, vol. 1, 51 - 54.
[2] KNIPŠILD', M. (1990): Kratkoe rukovodstvo k programme VERSTEU. Bjulleten' fonetičeskogo fonda russkogo jazyka, No.3, Bochum, 94 - 96.
[3] LEVINSON, St. (1983): Pragmatics. Cambridge.
[4] SAPPOK, Ch. (1990): Sintez replik kak modul' fonetičeskogo fonda russkogo jazyka. Bjulleten' fonetičeskogo fonda russkogo jazyka, No.3, Bochum, 5 - 39.

THE BEGINNING OF GERMANIC PROSODY

Anatoly Liberman

The University of Minnesota, U.S.A.

Early Germanic was a mora counting language; even after stress was fixed on the root, it could fall on either mora of a bimoric complex. In the northern dialects, two boundary signals also existed, the prototypes of stød (correption) and its opposite.

Very little progress has been made in the study of Germanic prosody since 1877, the year Verner published his article. All we know for certain about Germanic stress is still only Verner's Law. My attempt to eliminate word stress and reconstruct sentence stress at that period was misunderstood by my critics as an attack on Verner's Law itself [1]. To go beyond Verner, we can resort to facts of two types - accents in old manuscripts and prosodic phenomena in modern languages and dialects. The data obtained from even such conscientious spellers as Orm and Notker are hard to interpret. Modern accents also pose numerous difficulties, but at least they can be observed in the pronunciation of native speakers, and they display sufficient variety to justify an attempt at a reconstruction. I have spent the time between the appearance of my earlier works on this subject [2] and the present studying West Germanic (WG) accentology. Below I will state my conclusions in dogmatic form; detailed arguments and references will be given elsewhere.

Prosodic units that go 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 *Akzentumsprung* (a process responsible for the variation of the *éa*-*eá* 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 no-stød 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 phonematic basis and the (original) number of syllables: apocopated words are accented differently from nonapocated ones. In both areas, the basis is the older distributional factor, the only one that existed prior to apocope. Danish dialectologists regard stød as a late prosodeme. One of the implications of their theory is that Danish stød and WG correption are unrelated, which alone makes their views on the chronology of stød untenable.

The WG analogue of stød distinguishes between open and close vowels. According to the main Ripuarian pattern, correption occurs on the reflexes of the old open vowels /a: e: o:/ 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 nonmonothongized diphthongs are correpted when the word is dysyllabic or apocopated and when the postvocalic consonant is voiced. In disyllabic 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 voiced obstruent, so in *Hunde* but not in *Kante*.

The vowels /i: u:/ do not belong with /a: e: o:/ because in WG they were treated as diphthongal groups, namely, as /ij uw/, on a par with /an el or/, and so forth. Correption marked the end of the bimoric sonorous basis. All the early Germanic languages were mora counting, and stress, as evidenced by *Akzentumsprung*, could fall on either mora of a bimoric complex; correption separated the part of the word that served as the locus of shifting stress. In words with diphthongal groups, correption occurred only before a voiced obstruent because a voiceless obstruent marked the end of the prosodically active string by its voicelessness. Diphthongs were accented like diphthongal groups: when smoothed, they did not differ from the other long open vowels, and when preserved as units with two distinct elements, they joined /ij uw el ar/, etc.

In our classification of phonemes, we often try to discover whether Early Germanic obstruents were phonologically voiced/voiceless or strong/weak. 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 point of 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 or strength will 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 /ij/ and /uw/), as well as old monosyllables with a combinatory basis, had no correption before voiceless consonants, and it is not known how these words were pronounced. Two situations can be imagined. In some cases, noncorrepted 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 correption was itself an independent boundary signal within the framework of an equipollent opposition. If correption 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 preaspiration. The distribution of preaspiration 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 preaspiration is related to stød as sleeptoon is to stoottoon and that at one time preaspiration was the "lazy" opposite of stød.

A difficult problem confronts us in areas in which correption and "extension" are distributed according to the "mirror rule," as compared to the Ripuarian one: words with the reflexes of /a: e: o:/ and of smoothed diphthongs do not have correption, and in the other words it occurs before voiceless, not before voiced, obstruents. In most of these vernaculars, correption is phonetically weak, whereas the extending accent is prominent. The riddle of the "mirror rule" will remain insoluble if we keep looking on correption as the only thinkable marker of old bimoric bases. If, however, we accept the possibility of choice by old systems - [MM'] (two morae and correption) or [MM~] (two morae and a pause) - the Ripuarian rule 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 Ripuarian, the opposite of correption is "nothing," in Kleve, Arzbach, etc., the opposite of "extension" is a weak shadow of forceful correption.

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 an opposition in which ['] and [~] were equal partners should not be dismissed offhand. In the Rhein-Limburg area, accents occur only in conjunction with apocope, and apocope can be marked by either "extension" or correption. In the Scandinavian languages, stød (correption) never marks apocope, but in Low Franconian it regularly does so. Frings was wrong in denying a close tie between correption and circumflex. In old monosyllables with spontaneous bases, correption, indeed, has nothing to do with circumflex, but in apocopated words it is an analogue of a two-peaked accent.

Apocope endowed one boundary signal with a new role, and its yield increased. Our ideas of phonological relevance 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 phonemics [3] and even more evidently so in prosody. The two boundary signals would not have emerged if they had had no use, but becoming a marker of apocope enhanced the unit's visibility. From an acoustic point of view only "extension" resembles the circumflex of northern Saxon dialects, but any signal of apocope comes close to or merges with the circumflex of general phonetics, and it is no wonder that both "extension" and correption are often perceived as two-peaked: the boundary signal that became the marker of apocope changed its realization under the influence of its new function. Even if the original opposition ['] - [~] was equipollent, the loss of endings turned it into privative: one boundary signal was chosen as the accent of apocope and became the opposition's marked member and the most easily discernible prosodic shibboleth of the entire prosodic system. Frings carried his point too far when he insisted on the equal importance of correption 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.

In Danish, spontaneous and combinatory accentuation are seldom distinguished. Only in East Jutland does one come across elg with stød and høns without stød (diphthongal groups before a voiced and a voiceless obstruent respectively). It is more probable that Danish dialects simplified ancient diversity than that WG developed the juxtaposition of two spheres, but there could always have coexisted more and less complex systems. It seems that in the epoch following the fixing of stress on the root the Germanic languages of the North made use of two boundary signals (abrupt and smooth) dependent on the type of phonematic basis. These signals acquired greater importance when they came to be associated with apocope and when the number of syllables rules arose. No extant evidence points to the existence of correption (stød) and extension in all the Early Germanic dialects, and there is no bridge from them 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 no foundation in either phonetics or phonology. Akzentumsprung as the principle of ancient sentence stress and two boundary signals in a restricted area are all that we have.

[1] LIBERMAN, ANATOLY (1990). "The Phonetic Organization of Early Germanic." American Journal of Germanic Languages and Literatures 2, 1-22, and see the polemic in the subsequent issues of this journal.

[2] LIBERMAN, ANATOLY (1983). "Germanic Accentology. Volume 1. The Scandinavian Languages." Minneapolis: The University of Minnesota Press, and (1984) "Scandinavian Accentology from a Germanic Perspective." In: The Nordic Languages and Modern Linguistics 5. Arhus, 93-115.

[3] LIBERMAN, ANATOLY (1987). "Complementary Distribution as a Tool of Phonological Analysis. With a Note on the e Sounds in Old High German." General Linguistics 27, 173-88.

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 negative finding two alternative views have been proposed. Some scholars [2, 10] believe that languages are perceived as syllable- or stressed-timed because of phonological factors. Thus, in contrast to syllable-timed languages, stress-timed languages allow complex consonant clusters in coda position, and may reduce all vowels to schwa in unstressed position. Moreover, the addition or supression of schwa affects syllabification in the former (i.e., French) vs the latter language group. Phoneticians are however reluctant to abandon acoustic and articulatory timing measures. It seems now well established that there is no clearcut dichotomy between the two language types. Moreover rhythmic differences among languages probably reflect the contribution of several durational and spectral constraints [6].

In this paper I will look for phonetic correlates of timing organization in Catalan. In spite of Catalan being a Romance language, its phonological make-up does not fully accord 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 [ə] always behaves as a syllable 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 fa ("he says"). They were composed of one stressed syllable ([pa]) and zero, one or two preceding and/or following unstressed syllables ([pə]) (see Table I). 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 [ə]), 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 = .9, 1 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 Rc.

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 lenghthening

There is very scant evidence in support of the hypothesis that syllable timing organization is incompatible with final lengthening. Final lentghening has been reported to occur in French [5], Spanish and Japanese [7]. It does not show up however in Italian stressed syllables and vowels [12].

Final lenghtening 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, more so for unstressed vs stressed syllables, vowels and consonants [English: 9; Italian: 12], and for stressed and unstressed vowels vs consonants [French: 5].

Stress- 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% according to speaker) 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

(a) from 0 to 1 in all sequences (i.e., [pa] vs [papə], [pəpa] vs [pəpapə], [pəpəpa] vs [pəpəpapə]) for all speakers;

(b) from 1 to 2 in all sequences (i.e., [papə] vs [papəpə], [pəpapə] vs [pəpapəpə],[pəpəpapə]vs[pəpəpapəpə]) 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., [pə] vs [papə]) (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 [pəpa]).

Figure 2 illustrates anticipatory and carryover compression effects for stressed syllables according to speaker Re. The figure shows much less stressed syllable shortening (and thus much less stressed vowel shortening) in the 2 vs 1 than in the 0 vs 1 following syllables condition. In particular stressed vowels in proparoxytones 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 anticipatory than carryover compression effects and thus suggesting the existence of a left-dominant foot structure [Swedish: 8; 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 vs left-to-right 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 which are located two syllables away from the stressed syllable (i.e., word absolute initial 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 conforms 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 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 syllables duration within the word decreases in the progression final>initial>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 [ə] 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

[1] BRUCE, G. (1987) Rhythmic alternation in Swedish. In C.E. Elert, I. Johansson and E. Strangert, Nordic Prosody III, University of Umeå, Umeå, 31-41.

[2] DAUER, R. (1983) Stress-timing and syllable-timing reanalyzed. Journal of Phonetics, 11, 51-62.

[3] DUEZ, D. and Y. NISHINUMA, (1985-86) Le rhythme en français: Alternance des durées syllabiques, Travaux de l'Institut de Phonétique, Aix en Provence, 10, 155-169.

[4] FARNETANI, E. and S. KORI (1986) Effects of syllable and word structure on segmental durations of spoken Italian, Speech Communication, 5,17-34.

[5] FLETCHER, J. (1989) Prosodic aspects of French speech rhythm. Paper presented at the 117th ASA Meeting, Syracuse, U.S.A

[6] FOWLER, C. (1981) A relation between coarticulation and compensatory shortening. Phonetica, 38, 35-50.

[7] HOEQUIST, Ch. (1983) Syllable duration in stress-, syllable- and mora-timed languages, Phonetica, 40, 203-247.

[8] LINDBLOM, B. and K. RAPP (1971) Some temporal regularities of spoken Swedish, Papers from the Institute of Linguistics, University of Stockohlm, 21.

[9] NAKATANI, L.H., K.D. O'CONNOR and C.H. ASTON (1981) Prosodic aspects of American English speech rhythm, Phonetica, 38, 84-106.

[10] OS, E. den (1988) Rhythm and Tempo in Dutch and Italian. A contrastive study. PhD Dissertation, University of Utrecht.

[11] TOLEDO G. A. (1988) Compresión y ritmo en español, Revista Argentina de Lingüística, 4, 68-89.

[12] VAYRA, M. (1989) Slittamenti timbrici e variazioni di durata nel vocalismo dell'Italiano Standard. Paper presented at the NATO Advanced Study Institute on Speech Production and Speech Modelling, Bonas, France.

[13] VAYRA, M., AVESANI, C. and FOWLER, C.A. (1984) Patterns of temporal compression in spoken Italian, Proceedings of the Xth International Congress of Phonetic Sciences, M.P.R. Van den Broecke and A. Cohen (eds.), Dordrecht, 541-546.

TABLE I. List of nonsense words used in the experiment.

1. ['pa]
2. ['papə]
3. ['papəpə]
4. [pə 'pa]
5. [pə 'papə]
6. [pə 'papəpə]
7. [pəpə 'pa]
8. [pəpə 'papə]
9. [pəpə 'papəpə]

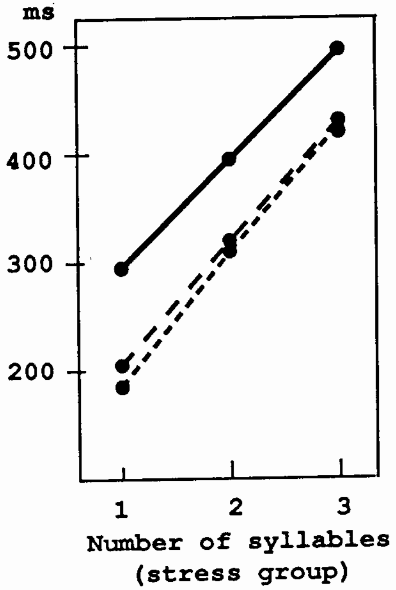


FIGURE 1. Stress group duration as a function of the number of syllables (speaker Re). The data are represented separately for oxytone (continuous line), paroxytone (dashed line) and proparoxytone (dotted line) nonsense words.

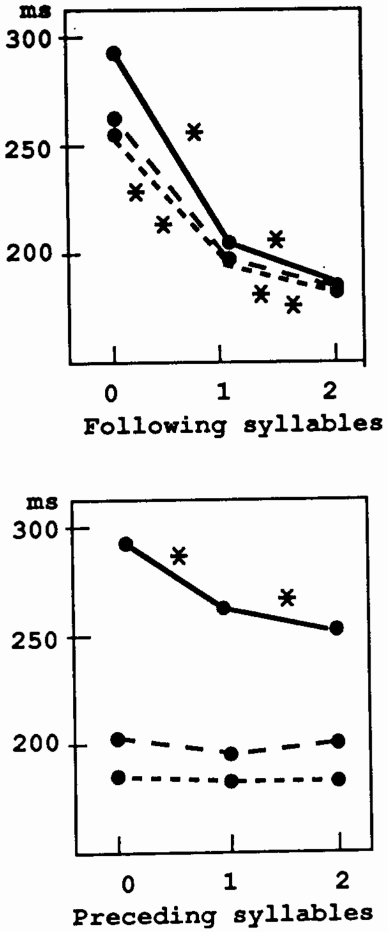


FIGURE 2. Anticipatory (upper graph) and carryover (lower graph) compression of stressed syllables as a function of the number of unstressed syllables in the stress group (speaker Re). The data are represented separately for one (continuous line), two (dashed line) and three (dotted line) preceding (anticipatory 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 interdependent, 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 would be expected to be susceptible to 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 constriction.

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 speakers' nose and upper incisor (to correct for head movement), lower incisor (to measure jaw movement), lower and upper lips, 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 data presented here are 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, disyllabic utterances of the form "mV_1CV_2", where V_1 and V_2 were /i/ or /ɑ/, and C was one of /p/, /pp/, /t/, /tt/, /m/, /mm/, /n/ or /nn/. In Japanese, utterances with /t/ or /tt/ as the intervocalic consonant and /i/ as the second-syllable vowel were excluded because /t/ palatalizes in this context. The Italian speaker produced 9 to 11 tokens of each utterance, and the Japanese speaker 12 to 16 tokens. The movement trajectories were digitized and smoothed prior to analysis.

Because of the very high correlations among the 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 5 ms frames, with the intention of extracting factors that would reflect the positioning of the tongue for the different 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.[1] 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 1. Trajectories measured.

Trajectory	Associated Gesture
Lower Lip vertical	initial /m/, bilabial intervocalic consonants
Tongue Tip vertical	alveolar intervocalic consonants
Tongue Dorsum horizontal	vowels
Tongue Body Rear vertical	vowels
Horizontal tongue factor	vowels
Vertical tongue factor	vowels only in Italian

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 /m/. In Italian (Figure 1), the large humps associated with the production of /i/ 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 /pp/ occurs earlier relative to the preceding lip movement ($p<.001$ for the effect of Length)[2] and to the tongue movement for the /i/ 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 /pp/ ($p<.001$ 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 neither 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

[1] Abbs, J., & Nadler, R. (1987), "User's Manual for the University of Wisconsin X-ray microbeam".

[2] Browman, C.P., & Goldstein, L. (1987), "Tiers in articulatory phonology, with some implications for casual speech", *Haskins Laboratories Status Report, 92*, 1-30.

[3] Browman, C.P., & Goldstein, L. (1988), "Some notes on syllable structure in articulatory phonology", *Phonetica, 45*, 140-155.

[4] Browman, C.P., & Goldstein, L. (1990), "Gestural specification using dynamically-defined articulatory structures", *Journal of Phonetics, 18*, 299-320.

[5] Fowler, C. (1981), "A relationship between coarticulation and compensatory shortening", *Phonetica, 38*, 35-50.

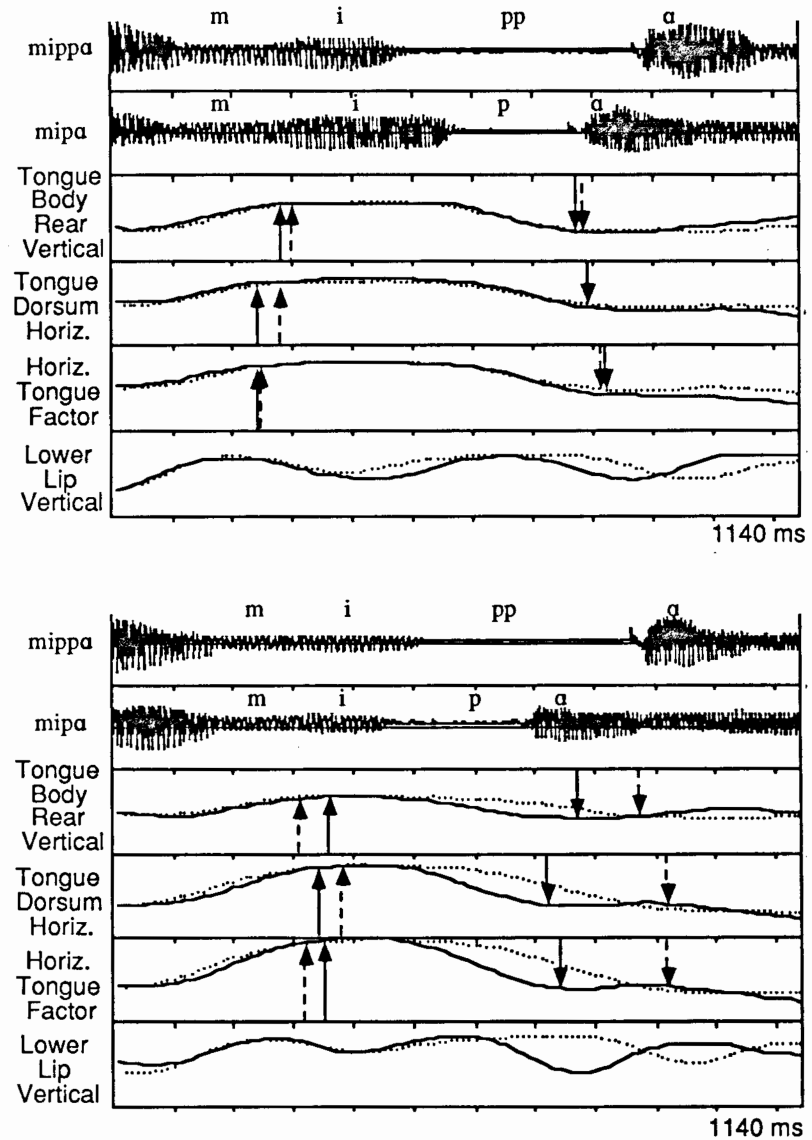
[6] Fowler, C. (1983), "Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet", *J. of Exp. Psych.: General, 112*, 386-412.

[7] Maddieson, I. (1985), "Phonetic cues to syllabification", in Fromkin, V., (ed.), *Phonetic Linguistics*, 203-221, Orlando: Academic.

[8] Öhman, S.E.G. (1966), "Coarticulation in VCV utterances: spectrographic measurements", *J. Acoust. Soc. Am., 39*, 151-168.

Work supported by NSF grant BNS-8820099 and NIH grant DC-00121 to Haskins Laboratories.

[1]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.
[2]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 /m/ 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 pausing 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 renderings 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 mingograms.

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.

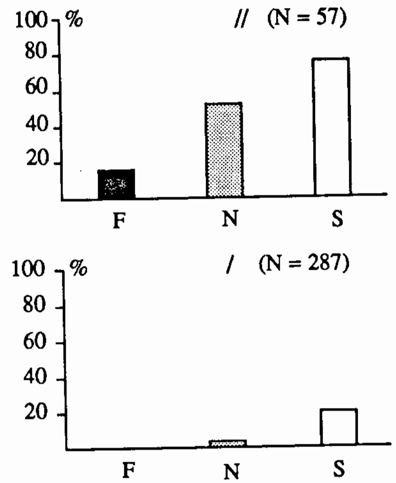


Figure 1. Relative frequency of strong pause positions at clause and phrase boundaries, in percent. Fast (F), normal (N), slow (S) rate. Based on 10 speakers.

A detailed analysis of the data indicates that different kinds of clauses did not attract pauses to the same degree. For example, complement clauses starting with *att* 'that' (as in *He said that* ...) were almost never preceded by a pause, not even at a slow speech rate. Temporal clauses, on the other hand, were generally preceded by a pause, as were conjoined clauses. Pauses also occurred very frequently before main clauses and some of the relative clauses. A similar pattern emerges from German data with comparable clause categories [2].

However, clauses of the same type were sometimes delimited by a pause, sometimes not. Length may be an important factor in these cases, as it seems that the probability of a pause between clauses is higher the longer and more complex the clause. Alternatively, it is not length but information load that is the important factor. Longer clauses and clause fragments contain more information than shorter ones. Thus, pausing may be a means for avoiding the clustering of too much information. That semantics plays a role for the insertion of pauses is supported also by the phrase data. The few phrase boundaries that were strong pause positions all delimited phrases with a high information load, viz., complex adverbial phrases and phrases expressing negation.

Thus, the present study suggests a multifactorial influence on pause distribution. (See [11, 13] for a more detailed account.) A similar complex basis for pausing is discussed by Umeda [15]. To isolate these determinants has the highest priority when it comes to predicting pausing. A number of studies have developed pausing algorithms as a means for revealing the "performance structures" of sentences [3, p 182-193; 6].

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 (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 is about .6 and at clause boundaries about .2 of the reference duration. In general the mean silent interval at phrase boundaries is somewhat

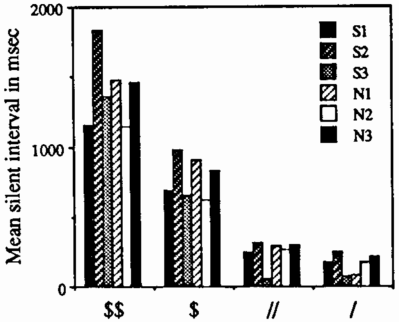


Figure 2. Mean silent intervals at paragraph, sentence, clause and phrase boundaries. Data for six speakers.

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 a sentence and a paragraph boundary [8]. Several studies indicate complementarity between lengthening and the following silent interval [4, 5].

5. PERCEPTUAL ASPECTS

The pauses in this study were aurally identified whereupon acoustic 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 to 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.) So far some preliminary observations have been made.

There is a high proportion of pauses without a silent interval. Over the six speakers the proportion ranges between 7 and 26%. In addition, there are many pauses with silent intervals 200 msec or shorter. Apparently there are other cues than silence to pause perception. Obvious candidates are Fo and intensity resetting, prepausal lengtening and voice quality irregularities. Several studies have shown that lengthening before a syntactic boundary may be a cue to boundary perception [7, 9]. Fo and intensity seem to be used as cues, too, but they are less effective than duration cues, including lengthening and silence [9, and references cited there]. A study of sentence and paragraph boundary perception points to a complex interaction of lengthening, voice quality irregularities (laryngealization), and silence [8].

Silence seems to be the more powerful cue. This may be inferred from the previous work cited above as well as from the present data. For example, listener agreement was 100% or close on pauses with silent intervals longer than 200 msec. It was the pauses with no or very short silent intervals (0-200 msec) that the listeners did not agree upon [12].

The silent interval, moreover, has to be adjusted to the specific boundary type. This conclusion may be drawn from a pilot experiment [16]. Three sections of the original recording of one speaker reading at his normal speed were stored digitally. The three sections each contained a boundary at which a pause had been perceived; one sentence boundary and two clause boundaries, one of which was longer than the other. In each section the boundary under study was preceded and followed, respectively, by a stretch of speech starting at the immediately preceding and following pause (boundary). A speech editing program made it possible for subjects to adjust the silent interval over a range from 0 to 1000 msec. The sections were tested one at a time and the

subjects alternated between setting a duration and listening to the result until they decided they had found the optimal silent interval. Each of the sections were tested three times in this way. The results are presented in Figure 3, which contains the original durations produced by the speaker alongside with the adjusted durations averaged over the three trials and nine subjects. Though the adjusted intervals are generally shorter than those originally produced, the temporal relations between the three boundaries are more or less the same in production and perception. These data, moreover, demonstrate that syntactic structure plays a role in the production as well as the perception of pauses.

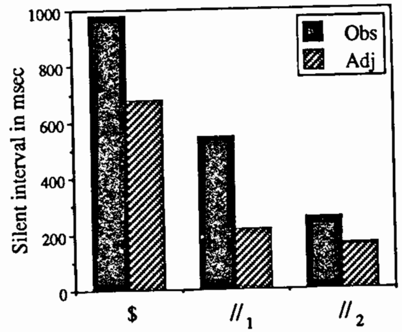


Figure 3. Originally produced silent intervals and adjusted intervals at one sentence and two clause boundaries. Averaged over 3 trials and 9 subjects.

6. REFERENCES

[1] BRINGMANN, E. (1990), "The distribution of Dutch reading pauses: A preliminary investigation on the influence of prosodic phrasing and punctuation on pause duration", Doctoral paper, Utrecht University.

[2] BUTCHER, A. (1981), "Aspects of the speech pause: Phonetic correlates and communicative functions", *Arbeitsberichte, 15,* Institut für Phonetik, Universität Kiel.

[3] COOPER, W. & PACCIA-COOPER, J. (1980), *"Syntax and speech"*, Cambridge, Mass: Harvard University Press.

[4] FANT, G. & KRUCKENBERG, A. (1989), "Preliminaries to the study of Swedish prose reading and reading style", *STL-QPSR, 2/1989.*

[5] FARNETANI, E. (1989), "Acoustic correlates of linguistic boundaries in Italian: A study on duration and fundamental frequency", *Eurospeech '89*, vol 2, pp 332-335.

[6] GEE, J. P. & GROSJEAN, F. (1983), "Performance structures: A psycholinguistic and linguistic appraisal", *Cognitive psychology, 15*, 411-458.

[7] KLATT, D. H. (1976), "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", *JASA, 59, 5*, 1208-1221.

[8] LEHISTE, I. (1979), "Perception of sentence and paragraph boundaries", In: Lindblom, B. & Öhman, S. E. G. (eds), *Frontiers of speech communication research*, New York: Academic press, pp 191-201.

[9] LEHISTE, I. (1980), "Phonetic manifestation of syntactic structure in English", *Annual bulletin, Research Institute of Logopedics and Phoniatrics, Tokyo, 14*, 1-27.

[10] ROOIJ DE, J. J. (1979), *"Speech punctuation. An acoustic and perceptual study of some aspects of speech prosody in Dutch"*, Dissertation, Rijksuniversiteit Utrecht.

[11] STRANGERT, E. (1990a) "Pauses, syntax, and prosody", In: Wiik, K. & Raimo, I. (eds.), *Nordic prosody V*, Phonetics, University of Turku, pp 294-305.

[12] STRANGERT, E. (1990b) "Perceived pauses, silent intervals, and syntactic boundaries", *PHONUM, 1*, 35-38, Department of Phonetics, University of Umeå.

[13] STRANGERT, E. (1991), "Where do pauses occur in speech read aloud?", *Proceedings from The Twelfth Scandinavian Conference of Linguistics*, June 14-16, 1990, Reykjavík, forthcoming.

[14] STRANGERT, E. & ZHI, M. (1989), "Pause patterns in Swedish: A project presentation and some data", *STL-QPSR, 1/1989*, 27-31.

[15] UMEDA, N. (1982), "Boundary: Perceptual and acoustic properties and syntactic and statistical determinants", In: *Speech and language: Advances in basic research and practice*, vol 7, New York: Academic Press, pp 333-371.

[16] Unpublished work in collaboration with Rolf Carlson and Björn Granström, KTH, Stockholm.

RHYTHMICAL STRUCTURES IN POETRY READING.

Anita Kruckenberg, 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 rhythmical continuity of mean interstress intervals. Similar rhythmical 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 expectations. Thus, a main difference between realizations of iambic and trochaic patterns is the relative shortness 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 perturbed 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 we refer to as phonetic feet.

The major problems of the present study are the following:

(1) To what extent is isochrony maintained in reading?

(2) How do pauses within and between lines maintain a rhythmical 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 syllabic 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.

As subject served a trained speaker, ÅJ, a language expert of the Swedish Radio, and a few of our laboratory staff. All of

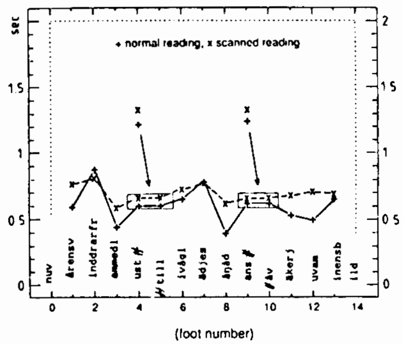


Fig.1 Interstress intervals in a "lab poem". Normal reading, +, and scanned reading, x.

them 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 line junctions to determine a possible rhythmical coherence with mean interstress durations.

In order to study durational correlates of specific meter types we performed a segmentation in terms of conventional syllables labelled as weak, W, and strong, S. An iambic foot thus contains a sequence of W+S and a trochaic foot the 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. We have also tested the consequences of relating the S and W component of a metrical foot, not to proper syllables but to vowel-to-vowel units with segment boundaries at stressed as well as unstressed vowel onsets. However, although such an analysis gave similar results it was discarded as a less reliable basis for the study of rhythmical structures.

3. RHYTHMICAL COHERENCE

We shall here report on some of the main results. A more detailed account will be given in [2]. The issues of isochrony and rhythmical coherence of pauses are illustrated in Fig.1, which pertains to a three-line iambic "lab poem":

"Nu vårens vind drar fram med lust,
till liv och glädje sång och dans
och väcker ljuva minnens 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 segmented at 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 thus preserves some degree of isochrony 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 in one 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 spanning feet have been 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 rhythmical continuation.

The same tendency of rhythmical 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 "Näcken" (The Water Sprite) by E.J.Stagnelius and the iambic "Kung Karl" (King Charles) by E.Tegnér. 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 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 an exact synchrony.

The iambic verse is made up of four line stanzas, normally with three complete (W+S) iambic feet per line with a regular occurrence of an extra weak syllable (hypercatalectic) 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 spanning interstress intervals to approximate 2Ta. In other words, at the end of the odd lines the pause acts as a supplement to the hypercatalex, 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 approximate a measure of nTa. This seems to hold rather well for rhythmical reading of prose. In poetry, on the other hand, we found a more consistent trend of the entire spanning interstress intervals to comply with a measure of mTo. This is what we could expect from a higher demand for rhythmical regularity in poetry reading. In a situation where To=Ta and both models fit the data, we can expect m=n+1, i.e. one rhythmical unit is contained in the physical sound segments of the spanning foot.

4. METER SPECIFIC PATTERNS

Our next problem has to do with rhythmical patterns of read poetry in relation to metrical patterns. If the W of the iamb does not differ significantly from the W of the trochee, and the same 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 syllable and also how 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 upbeat syllable, "anacrusis", added. However, these readings merely demonstrate that consistent, meter specific patterns could be produced, but they do not have a general proof value concerning unbiased performance.

In the reading of true poetry the situation was different. The subject was a trained reader but lacked any preconcept of a specific meter implied rhythm. However, the results from these readings supported our expectancy. We found that the duration 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 approximately 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, i.e. 35 ms, is thus 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 iambic S than the trochaic 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 of an 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, the W in a trochee and the S in an iamb. This would especially be the case of a terminal lengthening at 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 dactyls show a relative predominance of disyllabic words. It remains to be seen to what extent word inherent stress patterns condition durational patterns in poetry reading.

We plan to compare our data above with data from the same poems read as prose. Meanwhile, we gain some support from our earlier analysis of prose reading [3], where we have developed models of how stressed and unstressed syllable durations increase with the number of phonemes contained. If 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 trochaic verse. This simple prediction of how language structure might impose constraints on poetry reading supports what we have already seen, that the contrast between the iambic and the trochaic 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 rhythmical module close to an average free foot interval. In prose reading similar rhythmical 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 rhythmical patterns. The strong syllable appears to be of about the same length in iambic and trochaic verse, whilst the weak syllable is significantly longer in trochaic verse than in iambic verse partly as a consequence of 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 within a metrical foot to comply 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 F0 and intensity variations into an overall model of poetical 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 Tercentenary Foundation, The Swedish Council for Research in the Humanities and Social Sciences and The Swedish Board for Technical Development.

REFERENCES

[1] Elert, C-C. (1970), Ljud och ord i svenskan. Almqvist & Wiksell, Stockholm

[2] Fant, G., Kruckenberg, A. and Nord, L. (1991), "Stress patterns and rhythm in the reading of prose and poetry with analogies to music performance", Contributions to MULASP, Music, Language, Speech, and Brain. Int. Wenner-Gren Symp., Stockholm 1990. Forthcoming. .

[3] Fant, G., Kruckenberg, A. and Nord, L. (1991), "Temporal organization and rhythm in Swedish", ICPhS 1991.

[4] Halle, M. and Keyser, S.J. (1971), English Stress. Its Form, its Growth and its Role in Verse.

[5] Kiparsky, P. (1975), Stress, Syntax and Meter. Language, 51, 576-616[6] Loots, M.E. (1980), Metrical Myths. An Experimental Phonetic Investigation into The Production and Perception of Metrical Speech. 's Gravenhage

[7] Newton, R.P. (1975), Trochaic and Iambic. Language and Style, No 8, 127-156

[8] Risberg, B. (1936), Den svenska versens teori, Norstedt & Söner, Stockholm

A CONTRASTIVE ANALYSIS OF SPANISH AND CATALAN RHYTHM

M. Carrió i 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 rhythmical categories traditionally clasiffied: in stress-timed languages speech rate increase shows a higher degree of reduction in unstressed than in stressed sylllables; 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 rhythmical 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 not 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, spectrograms 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 time (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 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 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).

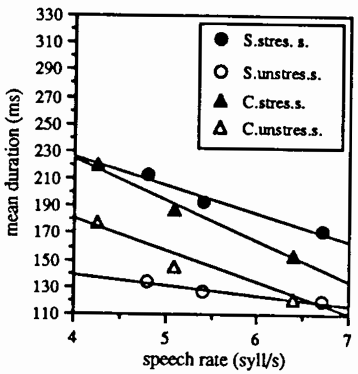
It is then clear that there are some problems connected with expressing speech rate in syllables per second. Questions arise as to whether pause-time has to be included and which types of syllables have to be counted, phonetic or linguistic ones. We have computed the overall values of linguistic and 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; for the same reason, values corresponding to speech rate have been used, because among other factores, 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 dependend 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:

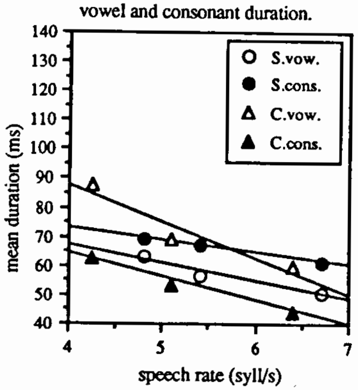
3.1. Syllables. Catalan unstressed and stressed syllables show an analogous shortening, which is higher in the stressed than in the unstressed ones (30.6 vs. 25.8). Spanish stressed syllables shorten to a lesser extend considering the behaviour Catalan syllables (20.4), and Spanish unstressed syllables present an even lower degree of shortening (7.2). See Figure 1.

Figure 1: speech rate vs. syllable duration.



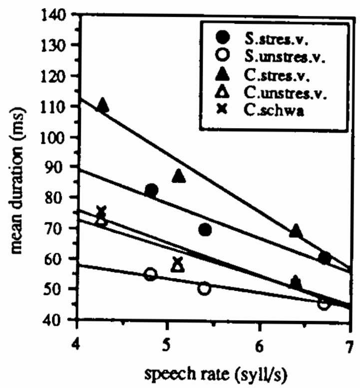
3.2. Vowels vs. Consonants. Vowels and consonants are subject to a similar shortening in both languages; however it is lower in Spanish (6.2 and 4.2) than in Catalan (12.6 and 8.7, respectively). See Figure 2.

Figure 2: speech rate vs. vowel and consonant duration.



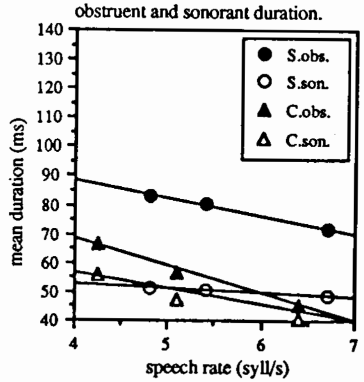
3.3. Stressed vowels vs. unstressed vowels. Differences in shortening between stressed and unstressed vowels in both languages are clear, although they are more prominent in Spanish (10.6 and 4.3) than in Catalan (18.2 and 9.0, respectively). Catalan has a schwa, which undergoes a shortening similar to the overall unstressed syllables (10.4). See Figure 3.

Figure 3: speech rate vs. vowel duration.



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 differences between sonorants and obstruents are not so important (9.6 and 7.0 respectively). See Figure 4.

Figure 4: speech rate vs. obstruent 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 reduction 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 preliminar results of this study suggest that more research is still needed in order to describe accurately the temporal compression phenomenon.

6. REFERENCES.

[1] BERTINETTO, P. (1981), "*Strutture Prosodiche dell'Italiano* ", Firenze, Academia della Crusca.
[2] CANTIN, M. & RIOS, A. (forthcoming), "Análisis experimental del ritmo de la lengua catalana", *Anuario del Seminario de Filología Vasca "Julio Urquijo"*, San Sebastián, Universidad del País Vasco.
[3] DAUER, R. M. (1983), "Stress-timing and syllable-timing reanalysed", *Journal of Phonetics, 11, 1* , 51-62.
[4] OS, E. den (1984), "Relations between tempo and duration of syllables and segments in Dutch and Italian", *PRIPU, 9,1* , 41-59.
[5] RIOS, A. (1991), *"Caracterización acústica del ritmo castellano"* , unpublished manuscrit, Laboratori de Fonètica Universitat Autònoma de Barcelona.

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 phonetical word, taking into account the inherent prosody of vocalic and diphthongal syllabic neuclei. The obtained model has revealed the main regularities in the distribution of duration of stressed and unstressed syllabic neuclei

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 vowels nor diphthongs in extended speech, we analysed vocalic and diphthongal syllabic neuclei 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 neuclei 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-syllable 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 neuclei makes 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 1st pretonic low vowels /a/ and /e/ (0.77:1 and 0.79:1 respectively);

b) there is a great difference in the length reduction of the 1st pretonic high vowels /u/ and /i/ (0.67:1 and 0.87:1 respectively);

c) the length reduction of the 2nd pretonic vowels is greater than that of the 1st pretonic vowels, /u/ being subjected to the highest degree of reduction (/a/,/e/,/u/,/i/ ⟶ 0.74:1, 0.76:1, 0.70:1, 0.81:1 respectively);

d) the length reduction of the 1st posttonic vowels is weaker than that of the 1st pretonic vowels (/a/, /e/,/u/,/i/⟶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 vowels is rather small, with /i/ being even longer in duration than the stressed one (/a/,/e/,/u/,/i/⟶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 1st pretonic long vowels is very similar to that of short vowels, with /ū/ being subjected to the highest degree of reduction. Long vowels /ā/ and /ē/ were not included into the experimental material as they are very rare in the pretonic position in the Lithuanian language. (/ė/,/ō/,/ū/,/ī/⟶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/,/ie/ and /uo/, 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 /ei/ has a greater length reduction in the 1st pretonic syllable than in the 1st posttonic syllable as in short and long vowels (0.71:1:0.74);

b) the diphthong /ie/ contrary to the diphthong /ei/ has a greater length reduction in the 1st posttonic syllable than in the

1st pretonic syllable (0.72:1:0.66).

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 (/áu/, /aũ/ — 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 1-st posttonic syllable (/ái/, /aĩ/ — 0.62:1:0.64, 0.77:1:0.80 respectively).

The analysis of durational ratio of stressed and unstressed vocalic and diphthongal syllabic neuclei revealed the temporal characteristics of a rhythmical model of a phonetical word. According to this model, the following regularities in the distribution of stressed and unstressed syllabic neuclei 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

/1/ ANUSIENE,L. (1987),"Duration of long stressed vowels in present-day Lithuanian utterances", Proceedings of XIth ICPhS,5,99-102

/2/ O'CONNOR,D. (1977),"Better English pronunciation" Cambridge Univ.Press.

/3/ PAKERYS,A. (1982),"Lietuvių bendrinės kalbos prozodija", Vilnius:Mokslas.

Incidences du trait phonologique de durée vocalique 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 régies 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, prosodiso-sémantiques et rythmiques que Mario Rossi (1985 et 1987) a formulées à partir d'exemples de français de Paris, sont communes au français des deux côtés de l'Atlantique.

Ces règles sont assez générales pour avoir un statut phonologique aux frontières des principaux constituants syntaxiques. Les intonèmes continuatifs ou conclusifs se retrouvent aux mêmes frontières au Québec et en France; c'est le cas des /CT/, /ct/, /CC/, /cc/ /paʀ/, même s'ils ne 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 débit, 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'Ile-de-France. La durée phonologique omniprésente dans le français québécois a des incidences sur les modes de réalisations phonétiques de l'accentuation, noeud du 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'énoncé. 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 ne le sont pas. Ces longues sont : /ɜ/ de fête opposé à la brève correspondante /ɛ/ de faites; le /ɑ/ de pâte opposé à la correspondante /a/ de patte; le /o/ de côte opposé à la brève /ɔ/ de cote; le /ø/ de jeûne opposé à la brève /œ/ de jeune; dans ce groupe de voyelles orales, l'opposition de durée s'ajoute à l'opposition de timbre et ne peut être neutralisée. Les quatre voyelles nasales sont aussi longues par nature (Santerre 1974).

Ces huit voyelles longues par nature s'allongent peu par coarticulation avec les constrictives sonores qui les entravent et elles se laissent peu abréger par les occlusives sourdes (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 /ɛ.a,ɔ,œ/ opposées aux quatre longues orales; deux voyelles, le /e/ et le /ə/, ne se trouvent pas en syllabe entravée.

La rencontre dans la rime des sept voyelles brèves avec les codas allongeantes, ou abrégeantes, ou indifférentes (occlusives sonores et constrictives sourdes) occasionne 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 ou allongées est considérable 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 [i] de comité peut faire de 0 à 5 ou 6 cs, tandis que le [ã] de commenter peut faire 12 à 20 cs. Ces rapports ne sont qu'indicatifs. Les voyelles hautes, en s'abrégeant ou en se syncopant en dehors de l'accent, abrègent et même font disparaître une syllabe, ce qui oblige les syllabes voisines à s'allonger pour intégrer les consonnes laissées sans noyau vocalique (Archambault 1985, J.-F. Couturier, recherche de doctorat en cours).

Durée morphologique lexicale

Les syllabes à noyau long par nature qui constituent des morphèmes lexicaux fréquents gardent leur durée vocalique 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 /tɜt/ et pâte /pɑt/ se prononcent en respectant la durée et l'entrave dans : tête à l'envers /tɜt a.../ et pâte à tarte /pɑt a.../ au lieu de /tɜ ta.../ et /pɑ ta.../. Entêté et empâté se prononcent /ã tɜ te/ et /ã pɑ te/ et jamais /ã tɛ te/ ni /ã te te/ ou /ã pa te / ni même avec un /ɑ/ abrégé. En québécois, les longues par nature conservent leur durée pertinente même en syllabe libre et en dehors de l'accent (Santerre 1990) [1]. 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: sage /saʒ/ a un noyau allongé qu'on ne trouve pas dans sagesse /saʒɛs/ à cause de la dérivation syllabique, mais qu'on retrouve dans sagement /saʒ mã/.

La durée dans la morphologie verbale

A la faveur des fusions vocaliques qui mettent en cause les flexions verbales, les contractions vocaliques sauvent les marques morphologiques de temps au moyen de la durée distinctive. Dans un test au moyen de phrases synthétisées, j'ai fait varier la durée vocalique dans la syllabe [ta] de /je ta po/ "Il est à Pau". Une centaine 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 Il était à Pau [je tɛa po] -> [je taː po]; quand on allonge le /e/ de /je/, on fait surgir au niveau phonologique la représentation des deux voyelles sous-jacentes de Il a été à Pau /jae tea po/ -> [jeː taː po], soit le passé composé. Ce test a été réussi presque sans exception par les Québécois, et n'a reçu que des réponses au hasard de tous les autres francophones présents (Santerre 1981).

Traces d'une ancienne durée

J'ai fait passer un autre test tout récemment sur la distinction de phrases "homophones" comme : "J'ai fait une partie d'échecs ce matin" et "J'ai fait une partie des chèques ce matin". Ces phrases lues par un locuteur parisien ont été complètement confondues par quinze auditeurs québécois; lues par un locuteur montréalais, elles ont été distinguées à 77%. Les mesures montrent que la durée des syllabes morphologiques autonomes 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 prosodie. 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 oeuvre comme en québécois une commande phonologique de production et de détection qui renforce l'effet mécanique involontaire.

Incidences 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 difficilement distinguées par les uns et par les autres. L'analyse prosodique instrumentale et psychoacoustique 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 faux sert à faucher l'oseille
0 2 0 0 0 0 0 1 (2-6)

(b) La faux sert à faucher l'oseille
0 0 2 0 0 0 0 1 (3-5)

Comme le prévoit l'auteur, (b) devient homophone de (c) : "la faussaire a fauché l'oseille".

Dans un test de compréhension auprès de seize Québécois étudiants de phonétique, (a) a été entendu comme la faux 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 faux parce qu'il accentue la syllabe suivante sert. La même phrase prononcée par des Québécois, qui déplacent aussi l'accent sur sert, ne change pas de sens, parce que faux conserve une durée qui sauve son statut de syntagme nominal sujet. L'accent de faux est fait par la durée et celui de sert est fait par l'intonation. Selon Dell, une règle de non-contiguïté accentuelle dans un même tronçon intonatif interdit d'accentuer en français deux syllabes consécutives. C'est sans doute parce que le larynx 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 n'empêche le locuteur de faire deux accents consécutifs pourvu qu'ils soient réalisés par des paramètres différents (Santerre 1990) [2].

La règle d'allongement de Dell

Dell (p.100) reconnaît que des phrases homophones comme (a) et (b) peuvent être distinguées par l'allongement d'une syllabe accentuée.

(a) Des dés odorants,
0 2: 0 0 1

(b) des déodorants
0 2 0 0 1

Cette règle stipule qu'on allonge facultativement la syllabe finale d'un mot accentuable, mais non pas la syllabe prépondérante des mots féminins qui est suivie d'un e muet, comme parle. C'est pourquoi Dell allonge l'accent secondaire dans (d) et non dans (c) :

(c)	ce -	lui	qui	par-	le	coud
	3	0	0	2	0	1
Fo:	153	151	120	176	135	112
Durée	154	149	137	257	102	200

(d)	ce -	lui	qui	part	le	coud
	3	0	0	2:	0	1
Fo :	149	154	140	185	126	124
Durée	143	157	153	336	112	202

L'écoute et les mesures révèlent que Dell accentue 2 par l'intonation seulement dans (c), et par l'intonation et la durée dans (d); 257 ms ne suffisent pas à faire sentir une durée ajoutée à la syllabe de trois phonèmes en (c).

En français québécois, la contrainte des mots féminins et celle de la non-contiguïté accentuelle me semblent respectées uniquement dans l'élocution très soignée de la lecture littéraire et du théâtre classique. C'est pourquoi un locuteur québécois peut réaliser couramment l'accent 2 dans (c) et (d) principalement au moyen de la durée et accessoirement au moyen de l'intonation.

(c)	ce -	lui	qui	parl'	coud
	0	3	0	2:	1
Fo:	124	156	(sourd)	139	105
Durée	136	211	150	369	240

(d)	ce -	lui	qui	par-	le	coud
	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 contiguïté accentuelle:

Celui qui part coud
0 3 0 2: 1

Dans la lecture de la phrase suivante, aucun locuteur québécois n'a fait entendre roucoulent, comme le fait Dell:
"Les seaux de l'élève roux coulent".

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 à une organisation temporelle 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

Archambault, D. (1985). *Production et perception de réductions de surface en français québécois.* Thèse de Ph.D. Université de Montréal.

Dell, F. (1984). "L'accentuation dans les phrases en français". Dell, Hirst et Vergnaud, *Forme sonore du langage.* Hermann, p. 65-122.

Rossi, M. (1985). "L'intonation et l'organisation de l'énoncé". Phonetica 42 : 135-153.

Rossi, M. (1987). "Peut-on prédire l'organisation prosodique du langage?"Etudes de linguistique appliquée, no 66 Didier Erudition.

Santerre, L. (1974). "Deux /E/ et deux /A/ phonologiques en français québécois". *Le français dans la région de Montréal.* Cahier de linguistique, no 4, p 117-145. Presses de l'Université du Québec.

Santerre, L. et J.-L. Chandon, (1981) "Duration distinguishes tenses in Montreal French". Actes du Symposium Prosodie, p. 28-41. Ph. Martin éd., University of Toronto.

Santerre, L. (1987)[1]. "Systématique des durées dans les rimes à voyelles longues et brèves par nature". Actes du XIe Congrès Intern. des Sciences Phonétiques. Tallinn, vol. 5, p. 126-129.

Santerre, L. (1987)[2] "Durées systématiques dans les rimes VC en fonction des segments et de l'accent". Actes des XVIe J.E. P., Société française d'acoustique, p. 229-232.

Santerre, L. (1990)[1]. "La désaccentuation des rimes à noyau bref ou long". Actes des XVIIIe J.E.P. Université de Montréal, p. 12-14.

Santerre, L. (1990)[2]. "La condition de non-contiguïté accentuelle en français; la théorie et la pratique". Revue québécoise de linguistique, vol. 19, p. 39-57.

PERCEIVING RHYTHM IN FRENCH?

Jacqueline Vaissière

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 tunes. PW layer is characterized 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 confusion 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 seem 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 realisations tend to correspond (see Figure 1).

1.1: The two BGs

The first ingredient is 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

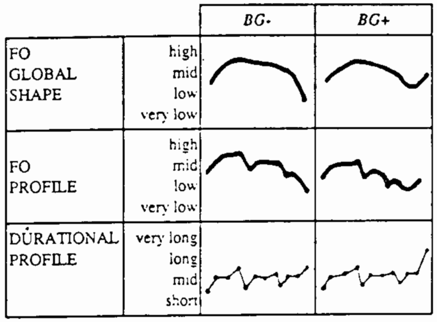


Figure 1: The two archetypal BGs and their decomposition into 3 PWs: the "valleys" in Fo and duration correspond to the function words, and the peaks to the final syllables of the PWs and BGs. The idealized curves correspond to BGs composed by 3 three-syllabic lexical words preceded by a monosyllabic function word.

Long BGs are prosodically substructured by an almost regular oscillation of the pitch between the high register and the low register and by durational contrasts. The reaching of the high register and a durational peak roughly corresponds to a PW in French (see Vaissière, 1983 for discussions on other languages).

1) long lexical word often corresponds to a single PW. Semantically related short words tend to be regrouped into a single PW. The tendency to have PWs of equal length in terms of number of syllables in read speech (probably as a consequence of rhythmic behavior) is so apparent in French that it was already included in an early model for speech synthesis (Vaissière, 1971).

2) the detailed Fo and duration profile of each PW depends mainly on the glide associated with the PW final syllable. As agreed by most phoneticians, the choice of a particular glide (rising with anticipation, rising, flat, falling) depends in part on the 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 (Vaissière, 1974:256).

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 acoustic realisation of the PW is obviously disturbed by a number of conflicting influences: (i) a short-long alternation (Duez & Nishinuma, 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 cointrinsic characteristics influence (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 Fo and duration profile from the idealized 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.

I.3: The CV syllables

The well-noticed perceptual saliency of the syllable as a rhythmic unit in French (Dauer, 1983, Wenk & Wioland, 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 intensity peak (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 (54.9%). 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 or less weight to one of the 3 main tendencies.

It is difficult to "prove" in a scientific way the coexistence of the different percepts in the speaker's mind. The "pregnant" speech patterns stored in speakers' memories are often said to influence the way they perceive the different languages. Delattre's examples of repetition of a sentence in a given language by natives of different languages (1965:23) seem to indicate that the stored basic patterns are very different (quite opposite) for French and English listeners. Results of psychoacoustic experiments on the perception of rhythm in non speech stimuli seem however to reveal that French and English archetypal speech patterns, apparently very different, may be 2 variants of the realisation of a universal pattern.

2: TEMPORAL VERSUS INTENSIVE RHYTHMITISATION

Psychoacoustic 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 accentuation (by increase intensity or pitch) as 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 and Allen, 1975). Perception of rhythm in speech seems to rely on the same basic principles as the perception of the rhythm in non speech stimuli.

The perception of intensity, pitch and duration in non speech stimuli (and in speech stimuli) are known to be not independent. For example, when some elements in a isochronous series are made more intense, the majority of listeners perceive the boundary before the accented burst ("rhythmitisation intensive", according to Fraisse, 1956, the listeners associating accentuation with beginning). Not all listeners react in the same way to the same stimuli. One third perceive the accented burst as group final ("organisation temporelle", the accented element is perceived as longer, and consequently as final). Fraisse therefore made the distinction between intensive rhythmitisation (relying on direct interpretation of increased intensity as right boundary marker) and temporal rhythmitisation (more intense elements seems to be lengthened elements, and therefore interpreted indirectly as final). Both rhythms may coexist in the same speech material, where more intense elements are often lengthened and their coexistence makes it more difficult to define rhythm in an easy way. In particular, it is difficult to estimate in some cases whether an accented element marks the beginning or the end of a rhythmic unit.

The inherent ambiguity between accentuation and induced lengthening may explain why French seems to avoid a strong accentuation of final syllables (because accented syllables tend to be perceived as initial), and overlengthening of non-final syllables in the group (because of the association between lengthening and right boundary). It also explains why emphatic stress falls on the word initial syllable, and not on the word final syllable. What makes interpretation of French rhythm more complicated is the fact that while the temporal organization (leading to the interpretation of the final syllable as the accented one) prevails, accentuation rhythm (like in English) marking the word beginning coexist in modern French.

Prominence on final syllables was generally considered to be the rule in non emphatic French. There is however a long series 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 while in conversational speech and story telling, final syllables were perceived as more prominent, in journalistic style, initial syllables were perceived as more prominent in 74% of the cases. The frequent regular 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 group may be interpreted as 2 variants of the same archetypal psychological pattern which associated accentuation with the beginning and lengthening with the end. In English, accentuation is dominant, and lengthening recessive. In French, temporal organisation is predominant, but (initial) accentuation is also intrinsically present (emphatic stress and initial rise), making the study of rhythm a very difficult matter. Progress may come from experiments in non speech stimuli and from investigation on how the same basic psychological constraints are integrated into the prosody and rhythm of diverse languages.

REFERENCES

ALLEN, G.D., (1975), "Speech rhythm: its relation to performance universals and articulatory timing", *J. Phon.*, 3, 75-86.

CARTON, F., (1967), "Pente et rupture mélodique en français régional du Nord", *VI Int. Cong. Phon. Sc.*, 237-241.

DAUER, R.M., (1983), "Stress-timing and syllable-timing reanalyzed?", *J. Phon.*, 51-62.

DELATTRE, P., (1965), COMPARING THE PHONETIC FEATURES OF ENGLISH, GERMAN, SPANISH AND FRENCH, Julius Gross Verlag.

DELATTRE, P., (1966), STUDIES IN FRENCH AND COMPARATIVE PHONETICS, Selected papers in French and in English, Mouton, London, The Hague, Paris.

DI CRISTO, A. & HIRST, D., (1986), "Modelling French microprosody: analysis and synthesis", Phonetica, 43, 1, 11-30.

DUEZ, D, & NISHINUMA, Y., (1984), "Some evidence of rhythmic patterns of spoken French", *PERILUS*, 1984-5, 30-40.

FONAGY, I, & FONAGY, J., (1976), "Prosodie professionnelle et changements prosodiques", *Le français moderne*, 44, 193-228.

FONAGY, I., (1980a), "L'accent français: accent probabilitaire (dynamique d'un changement prosodique", in L'ACCENT EN FRANÇAIS CONTEMPORAIN, *Studia Phonetica* 15, Fonagy, I. & Léon, P. (Eds.).

FRAISSE, P., (1956), LES STRUCTURES RYTHMIQUES, Louvain, Publications Universitaires.

FRAISSE, P., (1974), PSYCHOLOGIE DU RYTHME, Collection SUP, Presses Universitaires de France.

HEFFNER, R-M., S., (1950), GENERAL PHONETICS, The Univ. of Winconsin Press.

TOUATI, P., (1987), STRUCTURES PROSODIQUES DU SUEDOIS ET DU FRANCAIS, *Trav. de l'Institut de Linguistique de Lund*, Lund University Press.

VAISSIERE, J., (1971), CONTRIBUTION A LA SYNTHESE PAR REGLES DU FRANCAIS, Thèse de Troisième Cycle, Univ. des Sciences et Lettres, Grenoble 1971.

VAISSIERE, J., (1974), "On French Prosody", *Quarterly Progress Report*, Massachusetts Inst. of Technology, Res. Lab. of Electr., No 114, 1974, 212-223.

VAISSIERE, J., (1983), "Language-independent prosodic features", in PROSODY: MODELS AND MEASUREMENTS, A. Cutler, A. & R. Ladd, (eds.), Springer-Verlag, 53-66.

WENK, B.J. & WIOLAND, F., (1982), "Is French really syllable-timed?", *J. Phon.*, 10, 193-216.

TONE PRODUCTION IN STANDARD CHINESE:
EMG DATA AND COMMAND-RESPONSE MODELLING

P.A. Hallé

Laboratoire de Psychologie Expérimentale, Paris, France.

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: Fo 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 Fo 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 sytems, 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, Öhman [8] proposed the first to model Fo 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 Fo (for a more accurate account, see Boë [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. Öhman and Fujisaki both identified simple patterns attached to the production of pitch accent. We applied the same ideas to model Fo contours in Standard Chinese, proposing that qualitatively stable patterns of commands be 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: /yi2ge X zi4/ (a character X). Target syllables X belonged to minimal series sharing the same segmentals in the four tones: [ʃi], [pi], [mi], and [xu] (in Pinyin transcription, /yi/, /bi/, /mi/, and /hu/). Those segmentals were chosen in order to minimize SH contribution to supralaryngeal articulation (some SH activity related to tongue backing was expected for /hu/). 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 downdrift on the last syllable of breath groups.

Hooked wire EMG electrodes were inserted in the CT, Vocalis, and SH. Correct insertion was checked with various non-speech manoeuvres before and after the experiment, and periodically during its course. Subjects pronounced the 16 sentences (4 segmentals x 4 tones) at a normal speech rate, in 10 separate blocks.

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 fluctuations. One utterance per sentence, the closest to the mean with respect to the duration between line-up events, served for time scale 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 voiced 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 concentrated. 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 segment-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 on running speech: the longer a tone 2 syllable, 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": one onset step command followed by one offset step command of opposite amplitude. For Chinese, we call such pairs of commands "tone commands". We use both "positive" and "negative" tone commands: positive ones have an onset step command of positive amplitude and raise Fo, while 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 versus offset step commands, and to positive versus 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 an 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 rime; 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 analysed 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 segmentals. Fig. 1 shows CT and SH activities, together with tone commands obtained for the segmentals /mi4/. Similar results obtain for other segmentals. 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 rime. 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 developped within a muscle, whereas commands just indicate target Fo values. Contraction of the CT for example, produces a motive force f_c which tends to lengthen the vocal folds. The linear system approximation entails that f_c counteracts mechanical resistances to motion: inertia, frictions, 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, f_c must overshoot the target value corresponding to the high level static equilibrium. When this level is reached, f_c drops down to the target value, and eventually faints 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
[1] BOE,L.J., COTTET,O. & PARADIS,L. (1983), "Modélisation des évolutions de Fo", *Bull.Inst.Phon.Grenoble 12*, 48-65.
[2] FUJISAKI,H. & SUDO,H. (1971), "A model for the generation of fundamental frequency contours of Japanese word-accent", *J.A.S.A. 27*, 445-453.
[3] FUJISAKI,H., TATSUMI,M. & HIGUCHI,N. (1981), "Analysis of pitch control in singing", in K.N.STEVENS & M.HIRANO (eds.) *Vocal Fold Physiology*, Univ. of Tokyo Press, 347-362.

[4] FUJISAKI,H., HALLÉ,P. & LEI,H. (1987), "Application of Fo contour command-response model to Chinese tones", *Proc. ASJ fall meeting*, 197-198.
[5] HALLÉ,P.A., NIIMI,S., IMAIZUMI,S. & HIROSE,H. (1990), "Modern Standard Chinese 4 tones: EMG and Acoustic patterns revisited", *Ann.Bull.RILP 24*, 41-58.
[6] HOWIE,J.M. (1974), "On the domain of tone in Mandarin", *Phonetica 30*, 129-148.
[7] KRATOCHVIL,P. (1985), "Variable norms of tones in Beijing prosody", *CLAO 13 (2)*, 135-174.
[8] ÖHMAN,S. (1967), "Word and sentence intonation: A quantitative model", *KTH STL-QSPR 2-3*, .20-54.
[9] SAGART,L., HALLÉ,P. & de BOYSSON-BARDIES,B. (1987), "Modélisation des évolutions de Fo en Pékinois", *Rap.ATP 955316*, 20 pages.

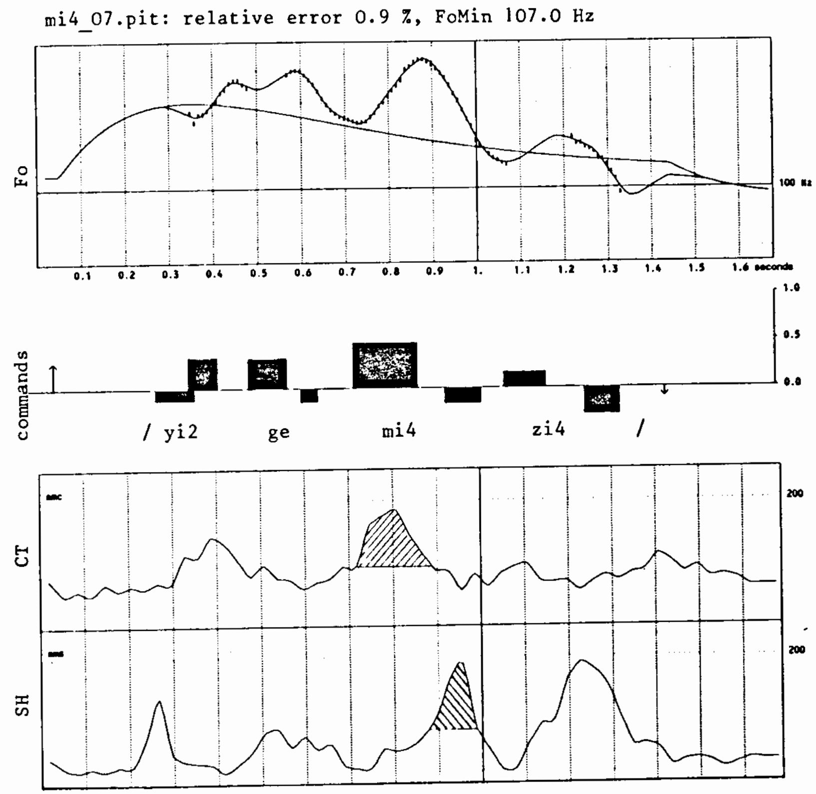


Figure 1. /mi4/: Actual and modelled F_0, model commands, CT and SH activities.

Table 1. EMG versus Commands: differences of timing (EMG-command, ms), ratios of amplitudes (EMG/command, arbitrary unit).

	Fo-raising			Fo-lowering		
	Δonset	Δoffset	ampl. ratio	Δonset	Δoffset	ampl. ratio
tone 1	+16 ms	+70 ms	3.78			
tone 2	+6 ms	+45 ms	3.63	-70 ms	-37 ms	11.2
tone 3				0 ms	-17 ms	12.6
tone 4	+25 ms	+52 ms	4.15	-21 ms	+18 ms	13.4

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 contexts like "Mississippi mud", has been attributed to rhythmic stress clash; the close approximation of two rhythmically prominent syllables is relieved by the left-ward 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 unacceptable, 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 legislator", but the secondary-stress-initial word in "Mississippi legislation" 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 indicate 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 not 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 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 syllable 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 mainstress syllables of the target word was judged unclear.

Of the 9 utterances of "Mississippi legislation", predicted not to undergo shift, 6 were shifted and 3 were unclear. Finally, of the 27 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. Nc 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 timing differences in the shifted-to syllable, the differences (as other investigators have reported) are not easy to demonstrate with this simple comparison between utterances.

F0 analysis: The F0 results present a somewhat clearer picture. We report here only the within-utterance measure of F0 change in the first syllable of the target words. This was defined as the size 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 those labelled 'unclear' were also associated with large rises (49, 34 and 16 Hz). All 3 of these cases were produced by the same speaker, and were instances where both the first and the mainstress syllables were strongly and equally prominent, suggesting that speakers can place pitch markers on more than one of the strong syllables of the target word under some circumstances.

The fact that stress shift was also perceived for a few utterances with no clear F0 change in the first syllable of the target word suggests that other acoustic cues may be used. Three of the five examples of this kind were produced by the same speaker, and there was no evidence that this speaker relied on duration increases: the initial syllable of the target word was 30-50 mS shorter in the stress-shifted utterances than in the corresponding single-word utterances. Other possibilities include a change in F0 from the last syllable of the preceding word, or the 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 shifted-to syllable 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 F0 patterns is the pitch marker observed in utterances of single words in frame sentences. An example is shown to the left in Figure 2, where the initial syllable "Mi-" shows an 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 overshadowed 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 Monaghan and Ladd [15].

Conclusion: The preliminary results reported here illustrate several significant points: (1) it is important to evaluate stress shift candidate utterances perceptually before measuring possible correlates of stress shift, since not all clash contexts invariably induce shift and it occurs in some non-clash 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, (3) 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; determining the options available to speakers will be an important step toward understanding the relation between not only rhythmic and intonational aspects of prosody, but also lexical and phrasal prominence.

Future work: Clearly, an understanding of stress shift will require a comprehensive approach involving phonological, acoustic-phonetic 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-shiftable pairs like "manageable" vs. "manageable legislators" will test this possibility.

4. ACKNOWLEDGEMENTS

This work was supported by grants NSF IRI-8805680 and NIH 8-301-DC00075. Conversations with P.J. Price and M. Ostendorf have been invaluable.

REFERENCES

[1] Abercrombie, D. (1965), Studies in Phonetics and Linguistics, London: Oxford U. Press

[2] Beckman, M. E. (1986), Stress and non-stress accent, Dordrecht:Foris

[3] Beckman, M.E. and Pierrehumbert, J. (1986), Intonational structure in Japanese and English, *Phonology Yearbook 3*, 255-309

[4] Beckman, M.E., Swora, M.G., Rauschenberg, J. anad deJong, K. , Stress shift, stress clash and polysyllabic shortening in a prosodically annotated discourse, presented at the Conference on Speech Processing, Kobe Japan, November 1990

[5] Bolinger, D. (1981), Two kinds of vowels, two kinds of rhythm, distributed by the Indiana University Linguistics Club, Bloomington

[6] Bollinger, D. (1986), Intonation and its parts, Stanford: Stanford U. Press

[7] Cooper, W. and Eady, S.J. (1986), Metrical phonology in speech production, *J. Memory and Language 25*, 369-384

[8] Gussenhoven, C. (1983), Stress shift and the nucleus, in On the grammar and semantics of sentence accents, Dordrecht: Foris (also in *Linguistics 21*)

[9] Halle, M. and Vergnaud, J.-R., An essay on stress, Cambridge, Mass: MIT Press

[10] Hart, J. 't and Collier, R. (1975), Integrating different levels of intonation analysis, *J. Phonetics* 3, 235-255

[11] Hayes, B. (1984), The phonology of rhythm in English, *Ling. Inq. 15*, 33-74

[12] Horne, M. (1988), Empirical evidence for a nonmovement analysis of the rhythm rule in English, *Lund Univ. Dept. Ling. Working Papers 33*, 139-152.

[13] Liberman, M. and Prince, A. (1977), On stress and linguistic rhythm, *Ling. Inq. 8*, 249-336

[14] Maeda, S. (1974), A characterization of fundamental frequency contours of speech, *Quart. Prog. Rep. MIT RLE 114*, 193-211

[15] Monaghan, A. and Ladd, D.R. (1987), An outline of the intonational component of the Edinburgh text-to-speech system, *Edinburgh Univ. Dept. of Ling. Work in Prog. 20*, 90-101

[16] dePijper, J.R. (1983), Modelling British intonation, Dordrecht: Foris

[17] Prince, A. (1983), Relating to the grid, *Ling. Inq. 14*, 19-100

[18] Nespor, M. and Vogel, I. (1986), Prosodic phonology, Dordrecht: Foris

[19] Pierrehumbert, J. (1980), The phonology and phonetics of English intonation, unpublished PhD dissertation, Massachusetts Institute of Technology

[20] Selkirk, E. (1984), Phonology and syntax, Cambridge, Mass: MIT Press

[21] Vanderslice, R. and Ladefoged, P. (1971), Binary suprasegmental features, *UCLA Working Papers in Phonetics 17*, 6-24

[22] Willems, N. (1982), English intonation from a Dutch point of view, Dordrecht: Foris

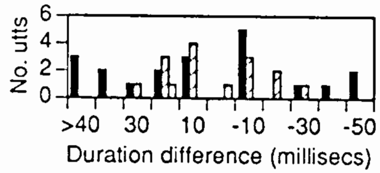
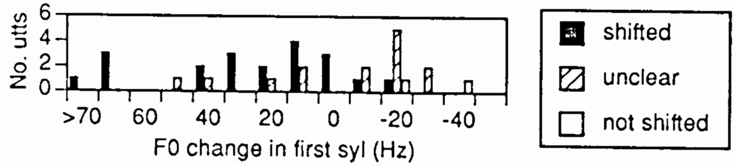


Fig. 1a (top): Difference in duration of the initial syllable for target words produced in a single-word phrase and in a corresponding stress shift candidate phrase by the same speaker

Fig 1b (bottom): F0 excursion in the initial syllable of target words produced in stress shift candidate phrases



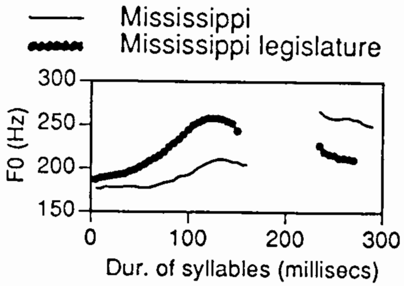


Fig. 2: Left traces are F0 values for the initial syllable Mi- in "Say the Mississippi again" (single word) and "Say the Mississippi legislature again" (stress shift candidate phrase) produced by a single speaker. Right traces are F0 values for mainstress syllable -sip- in the same utterances. Time axis reflects elapsed time for each syllable but not between syllables; onsets of voiced portions of syllables are aligned.

TERMINALITY AND COMPLETION IN DANISH, SWEDISH AND GERMAN

Nina Grønnum

Institute of General and Applied Linguistics University of Copenhagen

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 one 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 is qualitatively or quantitatively different from preceding ones, ceteris paribus. The difference may reside within the stressed syllable (a change in the magnitude of its Fo movement and/or in the direction of movement) and/or in the course of the post-tonic syllables. (2) The last stress group is positioned outside the range envelopping the preceding part of the utterance. -- (1) and (2) are not mutually exclusive. -- (3) The contour prior to the final event shows no principled variation with either utterance length or terminality. See (a).
- **Global** cues: (1) The final 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, Næstved, Aalborg, Tønder (Danish) and Malmö (Skanian)) 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 'globals').
- 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?

<u>(a) Utterances with final (or no) accent</u>

- Final falls in **Stockholm** are uncontroversially separate completion signals, tagged on to the sentence accent rise. The terminal vs. non-terminal cue lies in the preceding accented syllable, which is higher in non-terminals, cf. (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 through 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 in (f: solid line, g) share an overall slope which is less steep than in terminal utterances of comparable length and (g) furthermore 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 prosodic terminality. The level of this last stressed syllable follows from differences in global slopes.

<u>(b) Utterances with non-final accent</u>

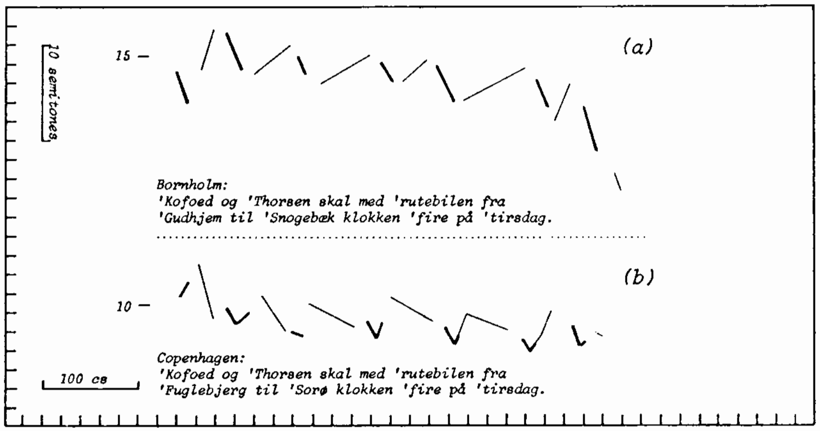
- If the highs and lows described above are indeed separate completion signals they must stay in place, at the end of the utterance, even if sentence accents and terminality cues move back. They should then be reached either progressively through or via a discontinuity in the preceding Fo course.
- **Stockholm** has only low completion cues and maintains an unmistakable low in final position: The post-accentual course can be regarded as a smooth interpolation between the early accent peak and the utterance final low, with diminished word accents superposed, cf. (h).
- In **Bornholm**, the final point in terminals constitutes the end of a generally smooth fall from the early accent, cf. (i). The fall from the high accented syllable in the **non**-terminal is not as deep and further movement is suspended until the final rise, cf. (j).
- In **German**, like in Bornholm, the initial accent is

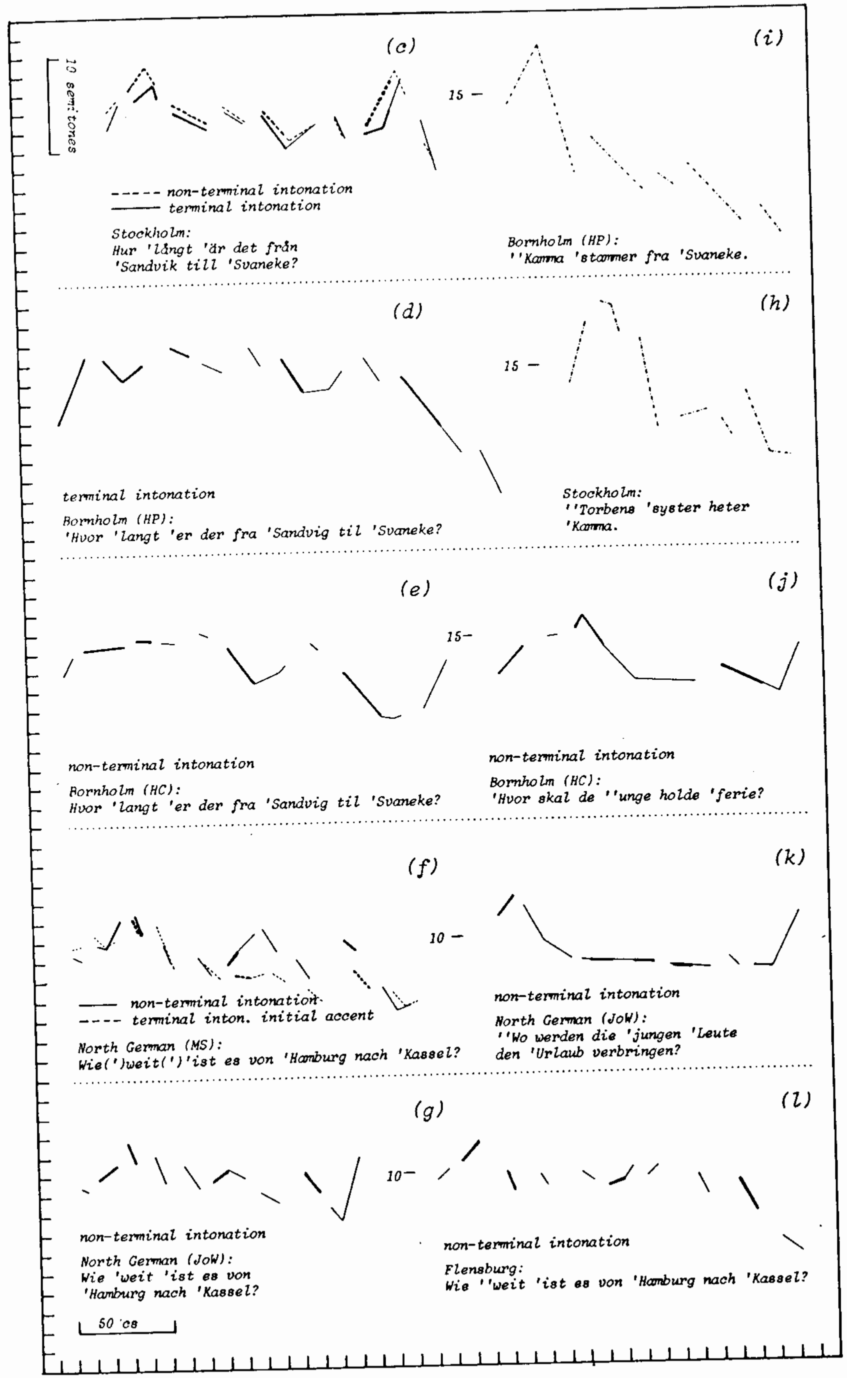
succeeded by a fall, which must be considered part of the accent command. In non-terminals, further movement is suspended, until the very final gesture, which may be either rising or falling, to the completion high or low, respectively, cf. (k, l). In terminals, the fall is continuous through the post-accented 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> Grønnum, N. (in print), "Prosodic parameters in a variety of Danish Standard languages, with a view towards Swedish and German", **Phonetica**.





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 accented syllables.
3. Association Domain (AD) boundaries, symbolized (...);
4. Scaling Domain (SD) boundaries, symbolized {...}. An SD contains a string of one or more ADs.
5. A phonological transcription of the intonation contour.

In (1) we give the representation of *Marjan* ('girl's name'), without the information under 5.

(1) {(m A r - *j 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: $\overset{*}{\mathrm{H}}$L, $\overset{*}{\mathrm{L}}$H, or $\overset{*}{\mathrm{H}}$LH. 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 SD.

Accentual Downstep. An AD may be provided with ACCENTUAL DOWNSTEP. This means that the $\overset{*}{\mathrm{H}}$ of each non-initial $\overset{*}{\mathrm{H}}$L will be realized with lower pitch than the preceding $\overset{*}{\mathrm{H}}$. ACCENTUAL DOWNSTEP is symbolized by placing the diacritic ! before the AD.

Accentual Downstep with Spreading. An AD which has been provided with DOWNSTEP, may additionally be provided with SPREAD. This means that each non-final $\overset{*}{\mathrm{H}}$ will be realized as a plateau instead of a peak, while the final $\overset{*}{\mathrm{H}}$ is rewritten $\overset{*}{\mathrm{L}}$, and the preceding L is deleted. Spreading-cum-downstepping is indicated by means of ~! before the AD.

Narration. NARRATION is applicable to $\overset{*}{\mathrm{H}}$L and $\overset{*}{\mathrm{L}}$H, and causes their $\overset{*}{\mathrm{T}}$ to spread. Thus, while an unnarrated realization of $\overset{*}{\mathrm{H}}$L 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 as the prefixation of a L tone segment, which is timed like the $\overset{*}{\mathrm{T}}$ of the undelayed Pitch Accent; it is symbolized by placing @ before the AD. The second modification, HALF COMPLETION, 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.

Phrasal Downstep. An SD may be provided with PHRASAL DOWNSTEP. This means that each non-initial AD will be realized with lower pitch than the preceding AD. PHRASAL DOWNSTEP is symbolized by placing ! before the SD.

Adjustment rules:

Boundary tones. Before the first Pitch Accent of an AD, an initial tone segment L is inserted, except when the first Pitch Accent is a narrated $\overset{*}{\mathrm{L}}$H, in which case H is inserted. This initial tone segment is referred to as %T. Moreover, $\overset{*}{\mathrm{H}}$LH rewrites as $\overset{*}{\mathrm{H}}$LH% in AD-final position. After the last Pitch Accent of the SD, L% is inserted after $\overset{*}{\mathrm{H}}$L, and H% after $\overset{*}{\mathrm{L}}$H, unless the AD has NARRATION or HALFCOMPLETION. The representations are transformed into 'narrower' transcriptions before they are processed by the timing and F0-rules. Domain-specific information is shown locally in these narrower representations. The adjustment rules change the representation in (2), for example, into (3). Observe that the boundary tone segments have been added, and that DOWNSTEP, SPREAD labels on the second AD have here been translated into the spreading of the first $\overset{*}{\mathrm{H}}$, and a rewriting of the second $\overset{*}{\mathrm{H}}$ as $\overset{*}{\mathrm{L}}$. Also, the DOWNSTEP label on the SD has been translated as the phrasal downstepping of the second AD.

(2) !{($\overset{*}{\mathrm{L}}$H $\overset{*}{\mathrm{L}}$H $\overset{*}{\mathrm{H}}$LH) ~!($\overset{*}{\mathrm{H}}$L $\overset{*}{\mathrm{H}}$LL%)}
(3) {(%L $\overset{*}{\mathrm{L}}$H $\overset{*}{\mathrm{L}}$H $\overset{*}{\mathrm{H}}$LH%) !(%L ~$\overset{*}{\mathrm{H}}$ $\overset{*}{\mathrm{L}}$LL%)}

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, $\overset{*}{\mathrm{T}}$, 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 $\overset{*}{\mathrm{T}}$ translate into two targets. All other tone segments translate into single targets. Where a tone segment yields two

```
(4)                     {( m     A     r  -   *j     A    n ))
 - temporal string:     123456 1234  123456  123456  1234  12345
 - tonal string:        %L ĤLH%
```

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.

- $\overset{*}{\mathrm{T}}$: locate Target at a distance of STARTIME of the vowel duration, counting from the beginning of the vowel. (Default STARTIME = 50%)

- ~$\overset{*}{\mathrm{T}}$: a spreading $\overset{*}{\mathrm{T}}$ (for SPREAD or NARRATION) has two targets:
Target1 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 preceding target. If the distance between preceding target and * is less than FROMTIME + TOTIME, locate Target midway between preceding target and *. The last T is located at TOTIME 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 FROMTIME after *.
Locate Target2 at TOTIME before the end of AD, if T% follows. If no T% follows, associate Target2 with last frame.

- T%: 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 (STARTIME). In (6b), we see the result of the other timing rules without OVERLAP. Target2 of %L 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 righthand boundary. Notice than the two targets overlap, as shown in the added target tier. Their associations are given as dotted lines to indicate their provisional status. Representation (6c) gives the state of affairs after the application of OVERLAP and SIMPLIFY. This representation is ready to go to the F0-rules.

```
(6) a. ((*j           A           n ))
       123456    12345678    12345678
           %L    H*L         H%

    b. ((*j           A           n ))
       123456    12345678    12345678
       T1 T2                 T1 T2
           %L    H*L         H%

    c. ((*j          A           n ))
       123456    12345678    12345678
       T1                    T1
           %L    H*          L   H%
```

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 H*) and a low reference value (which equals that of L*), together defining a register, whose width is is referred to as TRANGE (i.e. the distance between H* 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 Accentual Downstep factor **da** determines the lowering of H* targets in an AD with accentual downstep. The distance between L* and the most recent F0-value for a (downstepped or undownstepped) H*-target is referred to as !TRANGE. For targets after undownstepped H*, !TRANGE equals TRANGE.

A Phrasal Downstep factor **dp** determines the lowering of AD's in an SD with phrasal downstep.

For targets other that 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 Horne [4].

4,2. F0-rules

- %T: Target1: L = STARTSINK of TRANGE (Default STARTSINK=35%);
  H = H*
  Target2 = (STARTSLOPE)*Target1 (Default STARTSLOPE= .9)

- H* (= high reference)
  L* (= low reference)

- !H* F0 as given by the Accentual Downstep factor **da**.

- L in final Pitch Accent = L* (Target1 and Target2). 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 = 25%)

- L% = ENDSINK of TRANGE (Default ENDSINK = -10%)
  H% = previous Target + (ENDRISE of TRANGE) (Default ENDRISE = 30%).

4.3. The F0(m,n)-module

The implementation model F0(m,n) is given below. It calculates the F0-value for the *n*th T* in the *m*th AD.

$$F0(m,n) = Fr * N dp^{Sp*(m-1)} * W^{T} * da^{0.5*Sa*(1+T)*(n-1)}$$

Parameters:

Sp = +1 if Phrasal Downstep, 0 if not;
Sa = +1 if Accentual Downstep, 0 if not;
T = +1 for H*, and -1 for L*;

Fr = Reference line at the bottom of the speaker's range (default: 50 Hz for men and 100 Hz for women)

N = Defines the range, or the mean starting value above Fr (Default: 2.1)

W = Determines the distance between H* and L*. (Default: 1.6)

da = Downstep factor for downstepping H* targets within the AD. ("Accentual Downstep". Default: .80 if Sp = 1, and .70 if Sp = 0)

dp = Downstep factor for downstepping AD's in the SD. ("Phrasal Downstep". Default: .90).

5. INTERPOLATION

Interpolation between targets is by means of a 2nd order spline function. Future research involves the evaluation of different measures that can be taken if the time provided by the segmental string is insufficient to produce interpolations with slopes that remain within a pre-set speed limit. One measure might be UNDERSHOOT. Targets other than those provided by T* and T% would be undershot. Another approach would be to create more space by adjusting the position of *, thus creating more space (SHIFT). A third might be STRETCH, which would increase the time available by lengthening the segments concerned.

REFERENCES

[1] Berg, R. van den, Gussenhoven, C. & Rietveld, A.C.M. (1991), "Downstep in Dutch: Implications for a model", To appear in G. Docherty & D.R. Ladd (eds.) *Papers in Laboratory Phonology II*. Cambridge: Cambridge University Press.

[2] Gussenhoven, C. (1988), "Adequacy in Intonation Analysis: The Case of Dutch". In H. van der Hulst & N. Smith (eds) *Autosegmental Studies on Pitch Accent*. Dordrecht: Foris. 95-121.

[3] Gussenhoven, C. (1991), "Tone segments in the intonation of Dutch", In Th.F. Shannon & J.P. Snapper (eds.) *The Berkeley Conference on Dutch Linguistics 1989*. Lanham (MD): University Press of America.

[4] Horne, Merle A. (1988), "Towards a quantized, focus-based model for English sentence intonation",. *Lingua*, 75, 25-54.

[5] Ladd, D.R. (1987), "A phonological model of intonation for use in speech synthesis", In J. Laver & M. Jack (eds). *Proceedings of the European Conference on Speech Technology*. Vol. 2. Edinburgh: CEP Associates. 21-24.

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 convincingly 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. In our speech database project we have started to collect material from a variety of speakers, including professional as well as untrained speakers [5]. The structure of the database makes it possible to extract relevant information by simple search procedures. It is thus easy to retrieve information on the acoustic realization 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 quantify these by using e.g. scaling 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 and trained listener groups.

An interesting alternative is to ask the listener to adjust presented 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. In a study by Öster & 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 not possible.

Note that the voice does not always give away the complete speaker attitude. It is often observed that misinterpretation of emotions occurs if the listener is perceiveing 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 Öster 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 synthetic matchings of a number 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 factors such as sighs, voice breaks and jitter, lip smacks, etc, which often contributed in a decisive way to the intended emotion. This means that a standard acoustic analysis of produced 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 applications. Nevertheless, 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, styles and emotions. Specifically, we investigated an interactive production paradigm. We asked subjects to sit at a computer terminal and change 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,

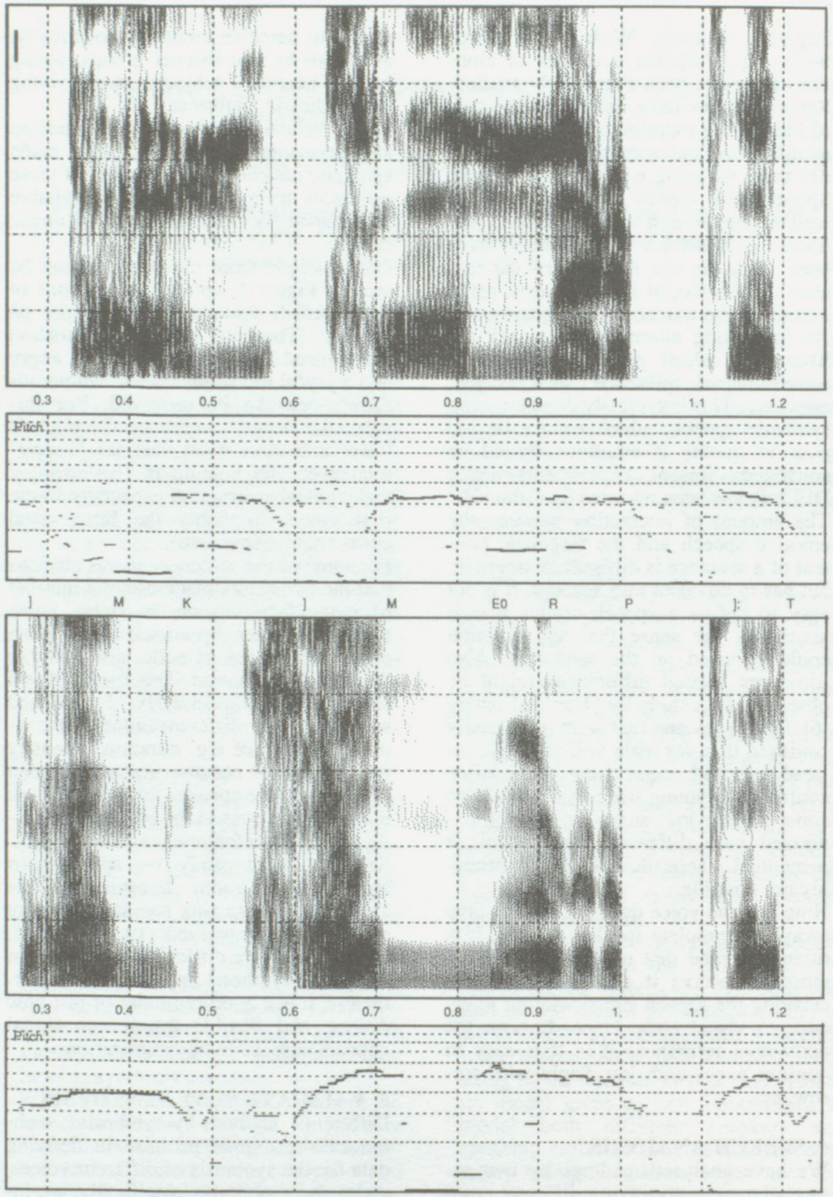


Figure 1: Spectrograms for two emotions imitated by an actor reading the sentence "De kommer på torsdag." /dɔm 'kɔmə r po 'tu:ʂda/. Only the underlined part is displayed Top: angry voice, Bottom: happy voice. In the pitch plot, horisontal dotted lines are 50 Hz apart starting at 100 Hz.

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 at 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 managable to the subjects, who responded with a fair degree of consistency. We are pursuing this line of experiments further including also voice source variations.

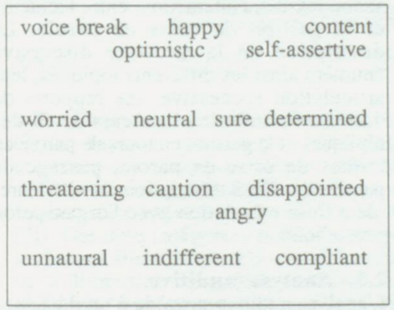


Figure 2. Example of free verbal responses in a speech synthesis production experiment with four subjects. See text for tested dimensions.

6. FINAL REMARKS

In this contribution we have indicated some ways of exploring speaker characteristics and speaking style using the speech database and synthesis environment at KTH. The work is still at a very preliminary stage. The presented example from emotive speech suggests that the described technique is useful also for other speech dimensions. Future applications of the gained knowledge are to be found in next generation speech synthesis and speech understanding systems.

ACKNOWLEDGEMENTS

This work has been supported by grants from The Swedish National Board for Technical Development, The Swedish Council for Research in the Humanities and Social Sciences, and the Swedish Telecom.

REFERENCES

[1] Bladon, A., Carlson, R., Granström, B., Hunnicutt, S. & Karlsson, I. (1987): "Text-to-speech system for British English, and issues of dialect and style", *European Conference on Speech Technology, vol. 1*, Edinburgh, Scotland.

[2] Cahn, J. E. (1990): "The generation of affect in synthesized speech", *Journal of the American Voice I/O Society, vol. 8*, pp. 1-19.

[3] Carlson, R., Granström, B. & Hunnicutt, S. (1990): "Multilingual text-to-speech development and applications", in A.W. Ainsworth (ed), *Advances in speech, hearing and language processing*, JAI Press, London

[4] Carlson, R., Granström, B. & Karlsson, I. (1990): "Experiments with voice modelling in speech synthesis", in Laver, J., Jack, M. & Gardiner, A. (eds.), *ESCA Workshop on Speaker Characterization in Speech Technology*, pp. 28-39, CSTR; Edinburgh.

[5] Carlson, R., Granström, G. & Nord, L. (1990): "The KTH speech database", *Speech Communication, vol. 9*, pp. 375-380.

[6] Öster, A-M. & Risberg, A. (1986): "The identification of the mood of a speaker by hearing impaired listeners", *STL-QPSR 4/1986*, pp. 79-90.

[7] Scherer, K. (1989): "Vocal correlates of emotion", in (Wagner, H. & Manstead, T., eds.), *Handbook of Psychophysiology: Emotion and Social Behavior*, pp. 165-197. Chichester: Wiley.

[8] Williams, C. E. & Stevens, K. N. (1972): "Emotions and speech: some acoustical correlates", *JASA vol. 52*, pp. 1238-1250.

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 de ce corpus à sa description en termes de variations des paramétres prosodiques (ici le paramètre de fréquence fondamentale ou Fo). Quatre analyses différentes sont ainsi appliquées à chaque corpus: une analyse discursive, une analyse auditive, une analyse acoustico-phonétique et une analyse par synthèse. Les données acquises à chaque étape sont représentées par une description discursive du corpus, une transcription prosodique sélective, des configurations tonales et des configurations tonales synthétisées par règles ou par LPC. Les deux questions majeures posées au cours de cette recherche sont d'une part celle de la relation entre la prosodie de la lecture en laboratoire et celle de la parole spontanée (pour ce qui est de la prosodie de la lecture en suédois et en français, cf. respectivement [1] et [6]) et d'autre part celle du rôle joué par la prosodie dans la structuration discursive de la parole spontanée. Précisons qu'avec la notion de 'parole spontanée', nous entendons un corpus produit et acquis hors de tout contrôle expérimental de la part du chercheur et dans des conditions de communication authentiques.

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 textuel, l'interaction entre locuteurs et la gestion de tours de parole. La description de la structure discursive énumère ainsi les différents topiques, leur articulation successive, les rapports de dominance entre les locuteurs au fil des répliques et la gestion en tours de parole en termes de prise de parole, passage de parole etc.. Cette description est ultérieurement mise en relation avec l'organisation prosodique.

2.2. Analyse auditive

L'analyse auditive procède à un décodage linguistico-prosodique du corpus. Elle se traduit tout d'abord par une transcription orthographique où sont indiqués les hésitations, les rires, les chevauchement de tours de parole etc.. (cette transcription est un préalable à l'analyse discursive). La transcription prosodique a essentiellement pour but de mettre en évidence la manière dont les fonctions démarcative et hiérarchique ont joué dans la structuration du corpus. Cette transcription est donc selective. Les cinq catégories selectionnées 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 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).

Tableau 1 (ci-dessous). Cinq catégories prosodiques avec leur transcription.

	Définition	Transcription
1) Proéminence accentuelle	Accent Focal ('Focal Accent')	''x
	Accent Primaire ('Primary Stress')	'x
	Accent Secondaire ('Secondary Stress')	,x
2) Regroupement prosodique	Frontière de groupe majeure	xx ‖ xx
	Frontière de groupe mineure	xx | xx
3) Registre de voix	Fortement étendu	↑ xx
	Légèrement étendu	↗ xx
	Normal	→ xx
	Légèrement réduit	↘ xx
	Fortement réduit	↓ xx
4) Marques de frontière	Frontière initiale avec ton montant	↗xx
	Frontière finale avec ton montant	xx↗
	Frontière initiale avec ton non-montant	non-marqué
	Frontière finale avec ton non-montant	non-marqué
5) Pauses	Pause courte	xx (.)
	Pause longue	xx (..)

2.3. Analyse acoustico-phonétique

Cette analyse procède 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 auditives, seules les pauses silencieuses relèvent clairement de la dimension temporelle du signal. La catégorie 'regroupement prosodique' délimite les domaines d'excercice 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 maxima et minima de Fo des représentations phonologiques intermédiaires en termes de segments tonals H(igh) ('Haut') et L(ow) ('Bas'). Les représentations phonologiques intermédiaires des accents du suédois et du français ainsi que leurs points de synchronisation syllabique 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 ['x]=H L*
accent II ['x̀]=H* L
accent focal
accent I [''x]=H L* H
accent II [''x̀]=H* L H

(2) Français

accent non-focal
['x]=L H*
['x]=H* L
['x]=L H*L
['x]= (D) L*
accent focal
[''x]= L H*

2.4. Analyse par synthèse

L'analyse par synthèse effectuée jusqu'à présent a eu pour objectif d'évaluer perceptuellement la valeur textuelle et interactionnelle de certaines configurations tonales (cf [4] et [7]).

3. EXEMPLES

3.1. Lecture versus spontané 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 confirment 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 ("'bien" et "'rir 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'accent 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 longs monologues textuellement hautement structurés et à choisir une manière de parler 'persuasive'. Certaines figures stylistiques caractéristiques apparaissent alors tels les intensificateurs, les parallélismes et les formes méta-discursives [5]. Les configurations tonales et leurs représentations phonologiques intermédiaires associées à ces figures stylistiques sont présentées dans les figures 4:1, 4:2 et 4:3. Un des politiciens étudié utilise par exemple des accents focals LH* avec un registre de voix étendu afin d'intensifier la valeur informative de son argumentation (cf. Fig. 4:2 et 4:3). Il atteint également une forme de parallélisme en répétant cette configuration combinée soit avec une frontière (cf. Fig. 4:2) soit avec une pause (cf. Fig. 4:3). Un parallélisme tonal est également produit par le maintien d'un registre de voix étendu sur plusieurs groupes prosodiques. Ces configurations tonales parallèles facilitent probablement la compréhension et la production de longs monologues en augmentant la redondance. En opposition à ce registre de voix étendu et ce mot à mot prosodique, un registre de voix réduit et un tempo plus accéléré est utilisé dans les commentaires méta-discursifs (cf par exemple la parenthèse du "↘ mais je "vois très 'bien | ↓ que vous ne le ferez 'pas ||" Fig. 4:1).

4. REFERENCES

[1] BRUCE, G. (1982), 'Developing the Swedish intonation model'. *Working Papers*, *22*, 51-116.

[2] BRUCE, G., WILLSTEDT, U., TOUATI, P. & BOTINIS, A. (1988), "Dialogue prosody", *Working Papers, 34*, 21-24.

[3] BRUCE, G. & TOUATI, P. (1990), "On the Analysis of Prosody in Spontaneous Dialogue", *Working Papers, 36*, 37-55.

[4] BRUCE, G., WILLSTEDT, U. & TOUATI, P. (1990), "On Swedish Interactive Prosody: Analysis and Synthesis", *Nordic Prosody V*, 36-48.

[5] NIR, R., (1988), "Electoral Rhetoric in Israel - The Televised Debates. A Study in Political Discourse", *Language Learning, 38:2*, 187-208.

[6] TOUATI, P., (1987), "*Structures prosodiques du suédois et du français*", Lund: Lund University Press.

[7] TOUATI, P. (1989), "De la prosodie française du dialogue. Rapport du projet KIPROS", *Working Papers 35*, 203-214.

5. ANNEXE

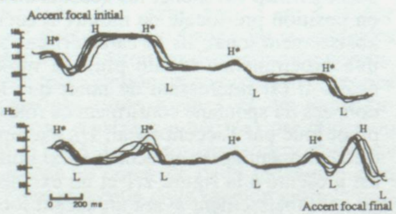


Figure 1. Abaissement et non-abaissement tonal en suédois lu (d'après Bruce 1982); l'effet d'un accent focal initial et final dans une phrase contenant quatre accents. Plusieurs configurations tonales produites par un même locuteur. Voir le texte pour des explications concernant les symboles (HL) utilisés dans la figure.

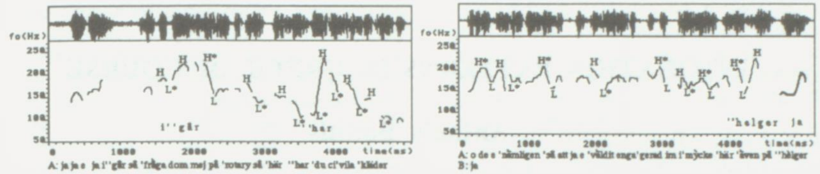
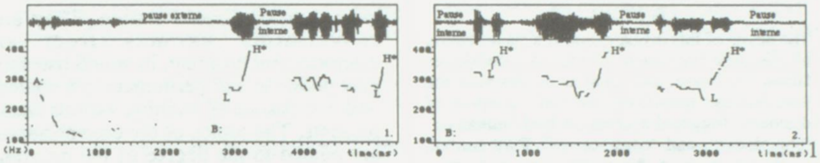
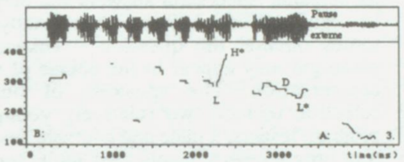


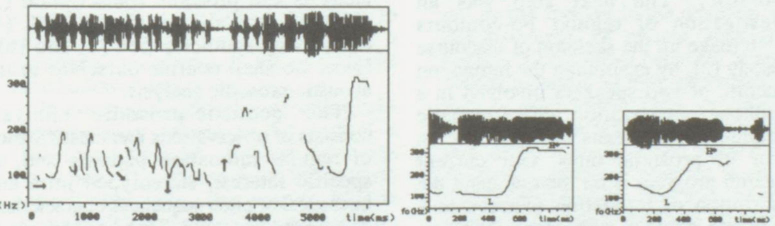
Figure 2:1 et 2:2. Abaissement (à gauche) et non-abaissement (à droite) tonal dans un dialogue spontané en suédois; effet d'un accent focal initial et final; onde sonore (en haut), configuration tonale (au centre) et transcription orthographique avec marqueurs prosodiques (en bas); chaque mot-clef est synchronisé avec un évènement tonal important; Voir § 2.3. pour des explications concernant les symboles (HL) utilisés dans la figure.





1. A: qu´est-ce que c´est la 'mort | d´après 'toi ||
   B: la "mort | (.) Y a des gens qui meurent 'pas (.)
2. B: s-sur'tout si (i)ls mangent 'bien | (.)
   peut pas mou'rir || (..)
3. B: quand on est très,très très très très "vieux |
   et ben on 'meurt ||
   A: ah 'oui ||

Figure 3:1, 3:2 and 3:3. Configurations tonales et représentations phonologiques de l'accentuation chez un enfant français (locuteur B); onde sonore (en haut), configuration tonale (au centre) et transcription orthographique avec marqueurs prosodiques (sur le côté); Voir § 2.2. er 2.3. pour des explications concernant les symboles (HL) utilisés dans la figure.



→une telle "loi à l'Assem'blée Natio'nale ||
↘ mais je "vois très 'bien |
↓ que vous ne le ferez 'pas ||
↑ je "crois que nous sommes le "seul pays au "monde ||

au "monde ||

où le "maire (.) est

Figure 4:1, 4:2 and 4:3 (de gauche à droite). Configurations tonales associées avec des figures stylistiques caractéristiques dans un débat politique français; 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 de transcription et (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 'continuative'. Furthermore, pitch range as a discourse turn and topic regulator 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 contrastive and interactive prosody [5]. We started our investigations by first analysing a face-to-face spontaneous conversation between two speakers, with optimal communicative conditions in the laboratory. The next step was an investigation of regular Fo-contours 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 (4) 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 regular intonation patterns and, of specific 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 contours

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 prosodic literature. By *stress* we mean prominent syllables with no reference to pitch whereas by *accent* we do mean pitch gestures whether they are co-ordinated with stress or not; in the former case, stress alternations make up the rhythm of the language whereas, in the latter, the interconnection of accents 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 one; the second phrase /ja na 'ðume 'tora/, apart from an initial pitch-gesture, has a widened pitch range, as a reinforcement of this (new) part of dialogue. 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 */'lipon/, i.e. a non-existent word in standard Greek. For this initial pitch gesture which, regardless of the rhythmic 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 ðer'matino xarto'filaka/ (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 /'ðiskolo/ 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 *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 intonative patterns (Fig. 3). The first two (3a, 3b) have falling final intonation but different communicative functions: (3a) is a pseudo-question, where the speaker is trying to win time or, in other cases in our data, to start or keep a conversation going; (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 intonative patterns, in which the final pitch gesture is co-ordinated with the final rather than the stressed syllable; these questions, the intonative pattern of which is very regular in our data, are 'emphatic' in the sense that the answer is of vital importance to the development of the discourse and, in this particular case, the outcome of the radio game.

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.

In an inter-speaker pitch contour adjustment, in certain environments, the turn-taking speaker's choice may be heavily dependent on the interlocutor's final pitch contour. Thus, an adult male finishes his phrase /ðila'ði 'nane kli'sto/ (4a) at a high pitch level and his interlocutor, an underaged male, responds with /ne/ at the same pitch level. Inter-speaker pitch adjustment, what we may refer to as *pitch-concord,* is evident also at a low pitch level; an adult female finishes her phrase /pine'lopi ki oδi'seas/ (Fig. 4b) at a low level and her interlocutor, an adult female also, responds /ve'veos/, 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.

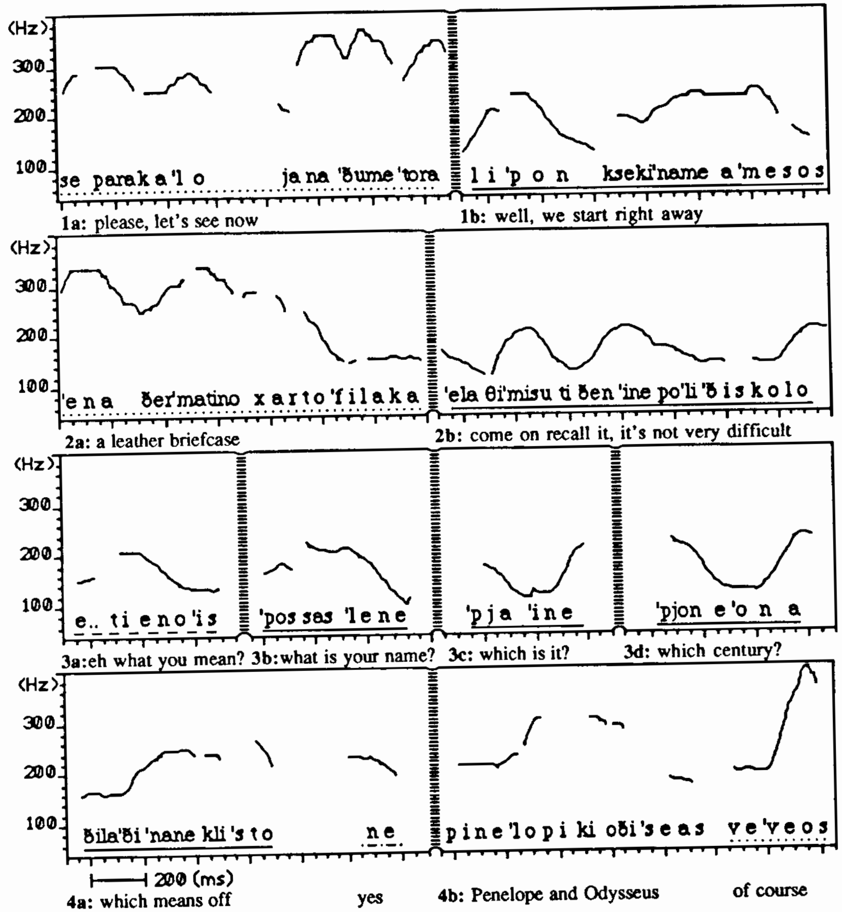
Although in a more comfortable conversation [5] we have found pitch-range as a turn and topic regulating discourse correlate, in the present material this phenomenon is drastically reduced. In other words, in a vivid interaction, speakers seem to take advantage of e.g. the presence of the completive accent or even the absence of a turn-keeping accent to intervene rather than using the pitch-range turn-leaving 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 regulations. 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 expansion directly reflects the involvement of the speaker(s) towards what it is said. It may span a succession of sub-turn-units and even have an inter-speaker effect. Pitch range may also indicate focus, although a (major) pitch-fall in combination with a post-focal accentless rhythmic organization is the rule as widely attested in Greek prosody. 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 discribed as 'focus' needs a 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 indiosyncratic 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 pitch 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 independently organized, with intonation as a *par excellence* TOP-DOWN prosodic parameter, specifically meaningful in interaction and discourse communication.



Figures: Pitch contours and pitch sequence extracts from different telephone conversations (see text). The full underlines represent a male program leader, the dots a female program leader, the dashed line an adult male program participant, the dots and dashes, a male child program participant, and no underlines, a female program participant.

BIBLIOGRAPHY

[1] Botinis, A. (1989a), *"Stress and Prosodic Structure in Greek"*, Lund: Lund University Press.

[2] Botinis, A. (1989b), "Discourse Intonation in Greek", *Working Papers* 35, 5-23. Dept. of Linguistics and Phonetics, Lund University.

[3] Botinis, A. (in press), "Greek Intonation", in Hirst, D. and A. Di Cristo (eds), *"Intonation Systems"*, Cambridge: Cambridge University press.

[4] Botinis, A. (forthc.), "The Organization of Intonation in Greek Discourse".

[5] Bruce, G., P. Touati, A. Botinis & U. Willstedt (1988), "Preliminary report from the KIPROS project", *Working Papers* 34, 23-50. Dept. of Linguistics and Phonetics, Lund University.

[6] Bruce, G. and P. Touati (1990), "On the Analysis of Prosody in Spontaneous Dialogue", *Working Papers* 36, 37-55. Dept. of Linguistics and Phonetics, Lund University.

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 rhythmical segmentation of the Spanish utterance. The differencies 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 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 countinuos speech segmentation into rhythmical 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 rhythmical units,among the sound means, the suprasegmental phenomena are almost exclusively taken into account.The rythmical units are, as a rule defined exclusively on the basis of only suprasegmental means conceived abstractely regadless of their concrete realizations in the flow of the speech, and on defining the rhythmical unit, exclusively one suprasegmental phenomenon is often taken into consideration. If the phenomena concerning the definition of the rhythmical unit delimitation are taken into account, only the pauses, which stand, in fact, outside the rhythmical unit itself, are considered. Therefore the so defined rhythmical 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 rhythmical unit as a rhythmical 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 speach potentially realized by a pause, by differences in the distribution of the positional variants of voiced consonatal phonemes, by glottal stop, by other sound phenomena or their combinations which, at the same time, carries suprasegmental means (stress, intonation, quantity) funcioning, in accordance with the role the rhythmical unit play 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 conceived rhythmical unit 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 can be described.

2.1. Rhythmical-semantic group delimitations

Important features differing Spanish from Czech are the differences in the distribution of variants 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 differ from one another in accordance with their position in the rhythmical semantic 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 Goya gója and de Goya [deɣója], Barcelona [barθelóna] and de Barcelona [deβarθelóna], etc. Czech native speakers do not respect this phenomenon and often pronounce the occlussive variants [b],[d],[g] in both positions.

Other phenomenon which is connected with the problem of delimitations of the rhythmical unit is the glottal stop. Considering that in Czech the glottal stop occures automatically at the beginning of utterance if the first phone is a vowel and the literary pronunciation requires the glottal stop after non-sylabic prepositions and in toher cases the pronunciation of the glottal stop is motivated phonostylically,Czech native speakers try to transfer their pronunciation with the glottal stop in all these positions into Spanish. Instead of en abir ena ríl they pronounce en a ríl .

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 distributio of the occlusive and the fricative variants of voiced consonantal phonemes.

When analyzing the syllable structure within the rhythmical unit in Spanish, we find that the sound coherence of the rhythmical-semantic group determines it division into syllables. It means that if a consonant, within the scope of the rhythmical-semantic group, ocures in an intervocalic position, it links to the next vowel and the syllabic division realites regardless of the boundaries of lexical units. The revore in the phrase han acabado the following syllabic [a/na/ka/βá/ðo].

Besidess the importance of the syllable as a component of the rhythmical-semantic group, we consider necessary to mention above all one of the features of the Spanish syllable: the tendency to

its openess. One of the manifestations of this phenomenon is after all the superiority of the syllable structure to the lexical one within the scope of the rhythmical-semantic group, as mentioned above, but also several assimilation phenomena become very important.
As for the type and the direction of assimilation, the articulation assimilation occures more frecuently, 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 [kontódo] - assimilation of the place of articulation, etc.
In the Czech language, the situation is rather different: the 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 these consonants with the assimilation of voice.

Other problem of the syllable structure of connected speech is closely relanted 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 pronunciation of expressions like a Ana [a.na] with the glottal stop [aʼana].

2.3. Stress

Further problems are linked up with the word stress within the scope of the rhythmical-semantic group. Unlike the Czech stress is fixed and has a delimitative function, Spanish is a language where the stress falls on different syllables, is considered as that of a giwen 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 can even be stressed syllable 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 considered constrastiv 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 other difference between Spanish and Czech consists inter alia in the fact that the quantity is phonological in the Czech language, while in Spanish the quantity (duration) is sometimes closely linked up with the stress position.
But considerring the quantity as a sound means of connected speech, we find that the relation between both studied lengages seems to be more complex. Changes of quantity (duration) in Spanish may be observed from two points of view: as a phono-syntactic phenomenon, i. e. as an 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 no phonological validity in Spanish often causes that Czech speakers do not respect differences in duration of 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 group 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 analyzis of an unadequate pronunciation of Spanish as foreing language, and it anables 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 in Czech, differences in assimilation, etc. is not sufficient.

5. REFERENCES

/1/ LARCOS LLORACH,E.(1974), "Fonología Espagnola", Madrid: Gredos.
/2/ DANEŠ, F. (1985), "Věta a text", Prague: Academia.
/3/ KULLOVÁ, J. (1988), "Mod ulaciones de la cadena hablada en español", Praga: AUC.
/3/ MALMBERG, B. (1965), "Estudios de fonética hispánica", Madrid:C.S.I.C.
/5/ PALKOVÁ, Z. (1974),"Rytmická výstavba prozaického textu", Prague: Academia.
/6/ QUILIS, A. (1981), "Fonética acústica de la lengua española", Madrid: Gredos.
/7/ ROMPORTL, M. (1973), "Studies in Phonetics", Prague: Academia.

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 "Sproeireën", 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. Hermes 1988).

* Guest researcher at the Institute of Perception Research, University of Technology, Eindhoven and guest co-worker at the Institute of Phonetic Sciences, University of Groningen as well.

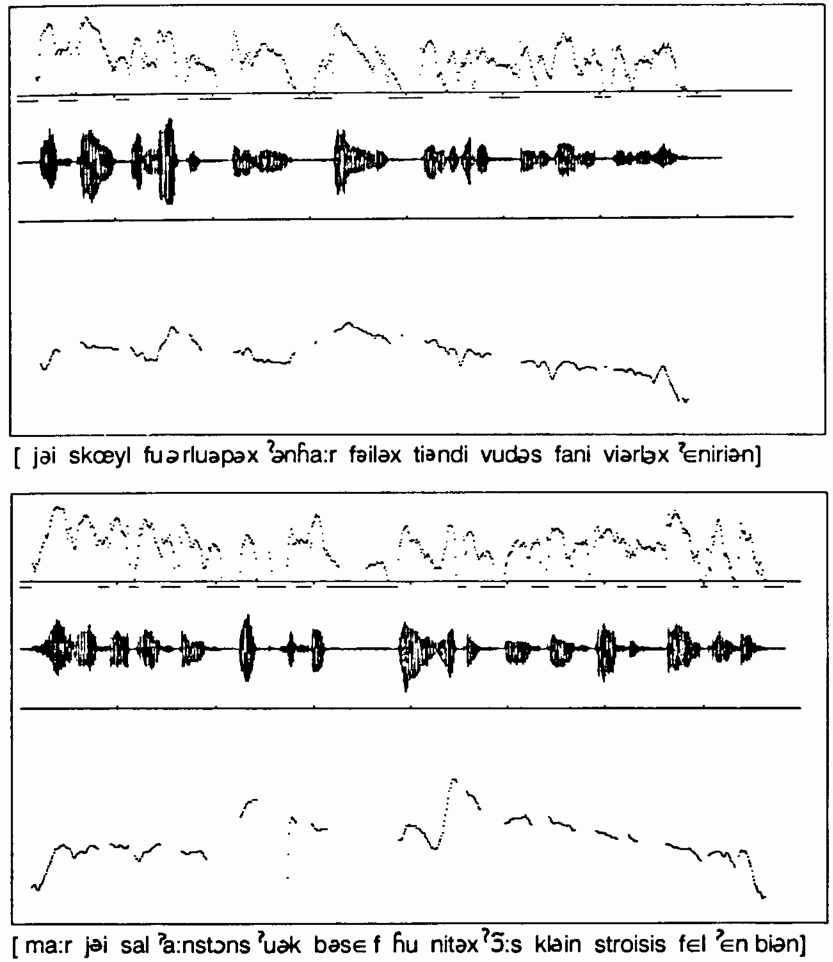
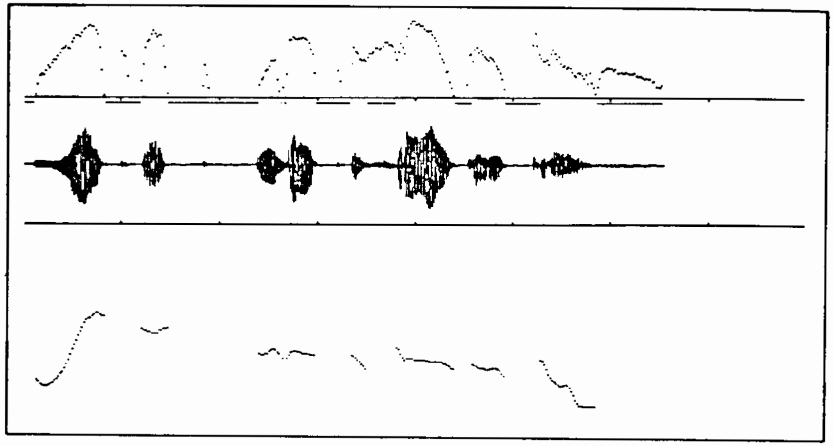


Fig. 1 The four lines of the poem "Skuiling" compressed into two run-on lines because of enjambment.

3. RESULTS

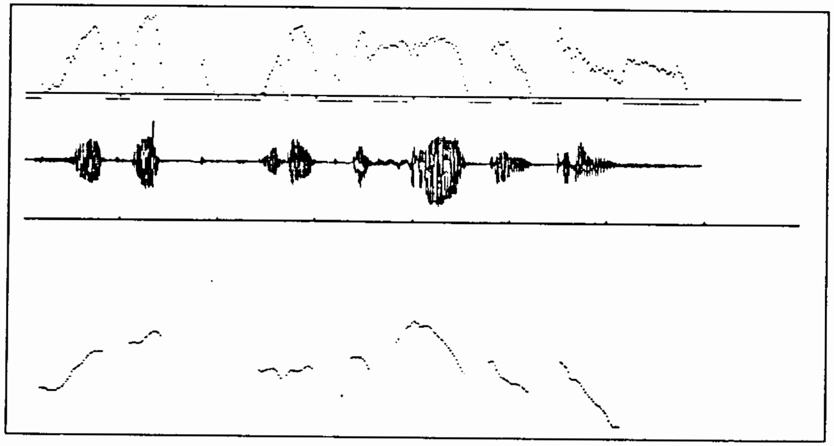
A print-out of the Fo contour of the four lines of poem no. 1 ("Skuiling"), clearly revealing which words are receiving prominence through increased pitch, is provided in Fig. 1.

Fig. 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"



[viət ˀɛk ɦudat ˀə frœu kan truəs]

Fig. 2 The realization of the line "...weet ek hoe dat 'n vrou kan troos" by a mother-tongue speaker.



[viət ˀɛk ɦudat ˀə frœu kan truəs]

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 "Skuiling" (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" ("safely") 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. Cruttenden, 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

[1] Cruttenden, A. (1986), "*Intonation*", Cambridge: Cambridge University Press.
[2] Hermes, D.J. (1988), "Measurement of pitch by subharmonic summation", *JASA 83*, 257-264
[3] Oakeshott-Taylor, J. (1984), "Factuality and Intonation", *Journal of Linguistics 20*, 1-21
[4] Scholtz, M. (1978), "*Die teken as teiken*", Kaapstad: Tafelberg.

EFFECTS OF VOICE CHARACTERISTICS ON ATTITUDE CHANGE

C. Gelinas-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's 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 Balloun, 1978); some other, more frequent, phonetic studies have investigated the mental image derived form 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 (of...) 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 counter attitudinal messages received daily would disengage from the exigencies of daily life. E.L.M. 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; Leon, 1971) or speaker's personality (e.g. Berger and Kellerman, 1989) or speaker's social status (Pittam and Gallois, 1986) or speaker's persuasive capacities (Brooke and Hung Ng, 1986) or arguments plausibility (Ekman, 1988). Hall (1980) 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 involvement 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 variable
- 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 (Zaichkowsky, 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 E.L.M.

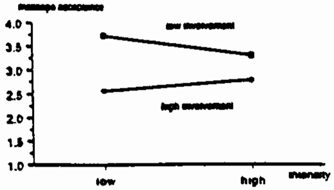


fig 1: combined effects of involvement and intensity

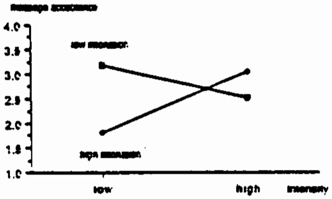


fig 3a: combined effects of intensity and intonation under low involvement

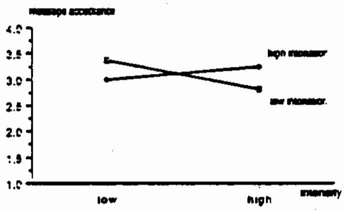


fig 4: combined effects of intensity and intonation

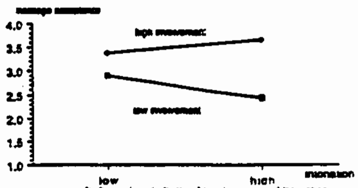


fig 2: combined effects of involvement and intonation on message acceptance

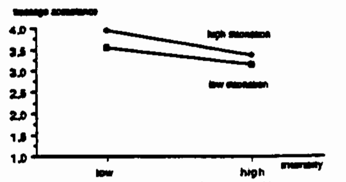


fig 3b: combined effects of intensity and intonation under high involvement

References

[1] BROOK, M.E. and NG, S.H., (1986, "Language and social influence in small conversational groups", *Journal of Language and Social Psychology*, 5 (3), 201-210.

[2] EKMAN, P., (1988), "Lying and nonverbal behavior: theorical issues and new findings", *Journal of Nonverbal Behavior, Special Issue: Deception*, 12(3,pt 1), 163-175.

[3] FONAGY, I., (1983), "*La vive voix: Essais de psycho-phonétique*", Paris, Payot, 346 p.

[4] GOLDBECK, T., F. Tolkmitt and Scherer, K.R.; (1988), "Experimental studies on vocal communication" in *Facets of emotion, recent research*, Scherer, K.R., ed., Hillsdale (New Jersey), Lawrence Erlbaum Associates Publishers, ch. 6, pp. 119-137

[5] HALL, J.A., (1980), "Voice tone and persuasion", *Journal of Personality and Social Psychology*, 38(6), 924-934.

[6] LÉON, P., (1971), "Essais de phonostylistique", *Studia Phonetica*, 4, 185 p.

[7] PAGE, R.A. and J.L. Balloun, (1978), "The effect of voice volume on the perception of personality", *Journal of Social Psychology*, 105, 65-72.

[8] PETTY, R.E., J.T. Cacioppo, and R. Golman, (1981), "Personal involvement as a determinant of argument-based persuasion", *Journal of Personality and Social Psychology*, 41, 847-855.

[9] PETTY, R.E., J.T. Cacioppo and D. Schumann, (1983), "Central and peripheral routes to advertising effectiveness: The moderating role of involvement", *Journal of Consumer Resarch*, 10, 135-146.

[10] PITTAM, J. and C. Gallois, (1986), "Predicting impressions of speakers form voice quality: acoustic and perceptual measures, *Journal of Language and Social Psychology*, 5 ($), 233-247.

[11] ZAICHKOWSKI, J.L. (1985), Measuring the involvement construct, *Journal of Consumer Research*, Vol. 12 (3), 341-352.

PAROLE CHANTEE ET PAROLE DECLAMEE: AUTOUR DE SALOME ASPECTS ARTICULATOIRES, RYTHMIQUES ET INTONATIFS

Sibylle Vater

Centre de phonétique expérimentale
Université de Genève, Suisse

ABSTRACT

There are a German and a French version of R.Strauss' Salome, and there is A.Mariotte's opera. This complex makes excellent material not only for comparing the vocal lines of two contrasting languages and relating them to Mariotte'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 phonémique et syllabique. Chaque compositeur s'y prendra à sa manière et en fonction de la langue du livret. Cependant, afin de se mettre au diapason du mot, le chant doit prêter à ce dernier ses propres structures, notamment celles d'ordre rythmique et tonal, celles donc que le linguiste qualifie de suprasegmentales.

2. SITUER LE SUJET

Notre étude cristallise autour du complexe de Salomé 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 désire rendre justice au texte français, ce qui l'engage à modifier la ligne vocale, l'orchestration restant inchangé. Pour réussir son entreprise, RS demande à R.Rolland(RR) des conseils en prosodie française [5]. De son côté, Jean de Marliave retraduit 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.Mariotte (AM) compose son opéra Salomé.

3. OBJECTIFS

Dans le cadre restreint du présent exposé, nous nous limitons à l'analyse de quelques exemples correspondants de
-la version allemande de RS(RSa),
-la version française de RS(RSf),
-Salomé d'AM.

D'une part nous comparerons la parole chantée avec la parole déclamée, d'autre part nous examinerons entre elles les lignes vocales des oeuvres impliquées. Plusieurs échantillons seront étudiés en fonction de l'articulation dramatique de l'opéra.

4. METHODE EXPERIMENTALE

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 dégagé de tout bruit technique éventuel. Nous visons des interprétations qui respectent l'écriture musicale dans son ensemble et nous remercions N.Jendly d'avoir bien voulu les assumer. Pourtant, qu'on ne nous tienne pas rigueur de l'exécution approximative de certains 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 abonde en symboles et phrases récurrents qui concourent à charpenter le drame et à créer un vertige, tout en se transmettant au besoin d'un personnage à un autre. RS et AM - ce dernier dans une moindre mesure - reprennent ces leitmotivs à leur tour, par exemple:
-RSf/AM Page: Vous la regardez toujours. Vous la regardez trop!
-RSf Salomé: Narraboth, je vous regarderai... N, regardez-moi!
(chez AM presque pareil)

Précisément, RS crée un "Musikdrama" où de nombreuses séquences sont intimement liées au resserrement de l'action. Trois escalades se profilent en particulier.
Nous en retiendrons ici celle qui entraîne l'exigence extrême de Salomé (scène 4).

6. PRESENTATION D'ANALYSES

6.1. L'exclamation initiale

Narraboth:
-RSa Wie schöon ist die Prinzessin Salome heute nacht!
-RSaf O ciel, combien la princesse Salomé est belle ce soir!
-RSf Comme la princesse Salomé est belle ce soir!
-AM Ah!... (suite comme RSf).

En passant de la diction à la déclamation rythmée et au chant, on constate d'emblée que la durée se multiplie par deux ou trois, voire plus, ce qui favorise avant tout l'épanouissement des noyaux syllabiques: Salomé, Nacht, belle, soir. Il en résulte que le caractère marginal des consonnes s'accentue. A leur tour, les conventions intonatives du parler, qui accordent une large part aux modulations coulissantes, sont abrogées: dans le chant les performances tonales sont essentiellement graduées, elles se constituent en tremplins et paliers. Mais associées à la parole, elles doivent se plier à des servitudes linguistiques. C'est pourquoi, par exemple, en adaptant la ligne vocale au texte français, RS abolit l'attaque aiguë (Wie schön...), incompatible avec une exclamation qui débute par comme. Cependant, cette modification contraint le compositeur à décaler la phrase entière d'une demi-mesure (v.ill.3, RSf). De notre côté, nous proposons une version française qui, à de minimes retouches de durées près, maintient la mélodie et les mesures.

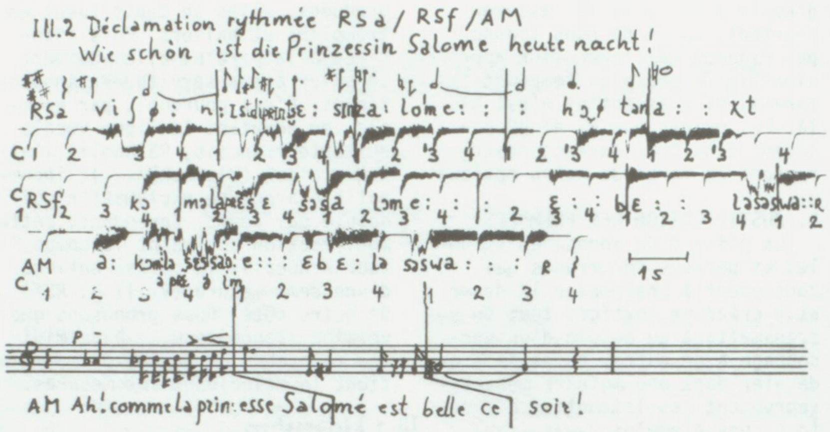
Ill.1 Déclamation

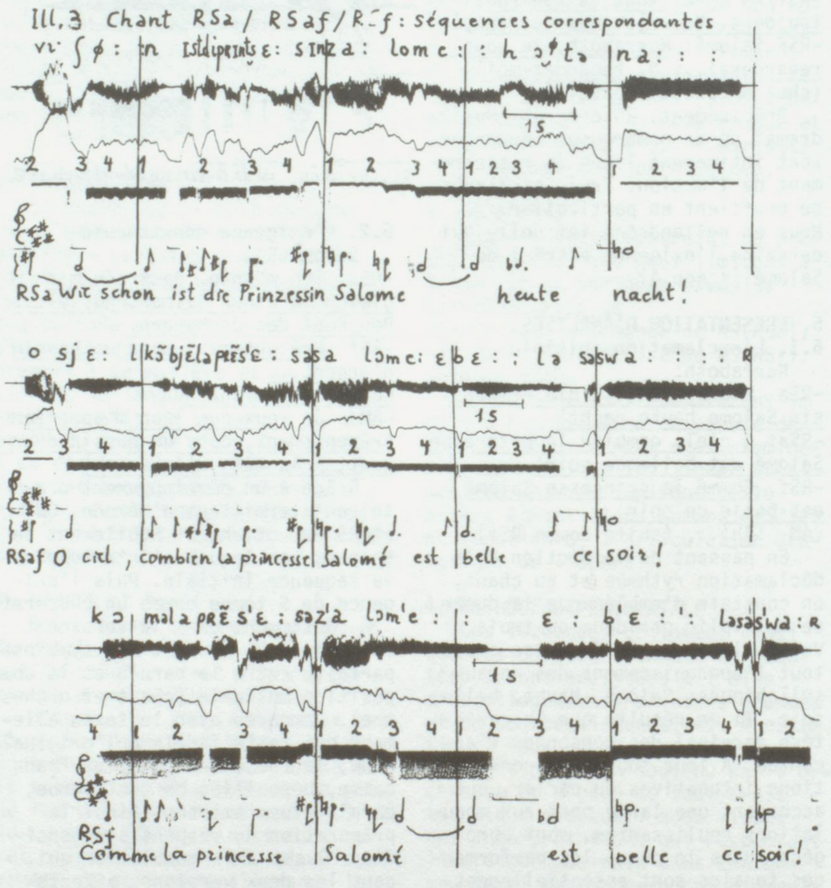
6.2. L'exigence monstrueuse

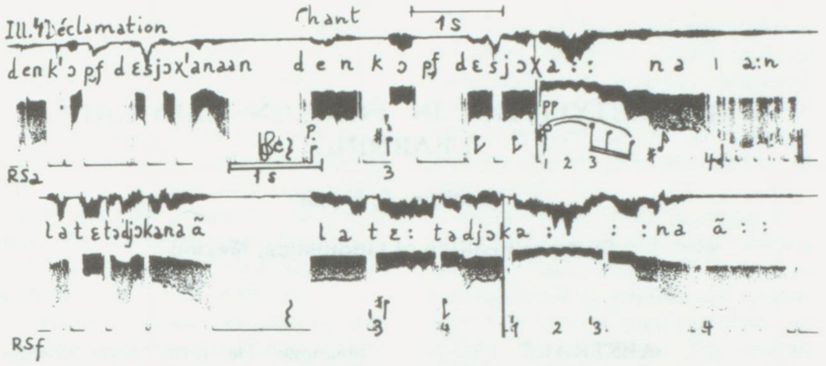
Salomé(S):
-RSa Ich möchte, dass sie mir gleich in einer Silberschüssel... Den Kopf des Jochanaan.
-RSf Présentement dans un bassin d'argent... La tête d'Jokanaan.
-AM Je veux que l'on m'apporte présentement, dans un bassin d'argent, - la tête d'Iokanaan!

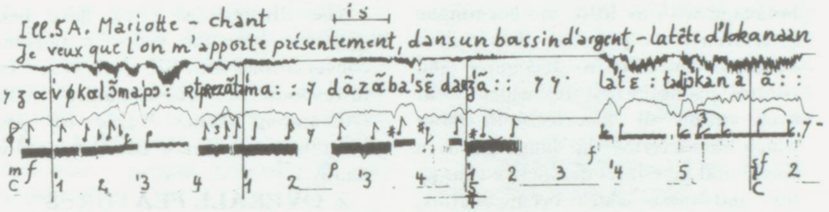
Grâce à un développement oratoire intermittent d'Hérode, OW et RS entretiennent habilement le suspens par lequel s'interrompt la séquence initiale. Puis l'exigence de S tombe comme un couperet.

Notamment dans la version française, RS réussit une symbiose parfaite entre la parole et la composition musicale (chant et orchestre). Comparée avec le texte allemand qui reste fidèle à l'original d'OW, la rhèse introductive française, dépouillée de tout verbe, paraît plus ramassée. Ainsi la préparation du suspens s'intensifie. Quant à la modulation qui, dans les deux versions, affecte









le nom du prophète, surtout le traitement rythmique, tonal et articulatoire du premier a démontre un éclatement de potentiel phonémique suprême (v.ill.4).

Au contraire, par la suppression de la parenthèse d'Hérode et par une ligne vocale plate, AM compromet gravement l'effet dramatique (v.ill.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.

8. REFERENCES

[1] BOISSIER,M.-N. et LIEVRE,G. (1990), "Salomé/O.Wilde,R.Strauss", Lyon: Opéra de Lyon.
[2] MARIOTTE,A.(1910), "Salomé", Paris: Enoch.
[3] STRAUSS,R.(1905), "Salome, in deutscher Übers. von H.Lachmann", Berlin: A.Fürstner.
[4] - (1906), "Salomé", Berlin: A.Fürstner.
[5] STRAUSS,R.et ROLLAND,R.(1959), "Correspondance. Fragments de journal", Cahiers R.Rolland, 3, Paris: A.Michel.

PHONOSTYLISTICS IN FOREIGN LANGUAGE LEARNING

Kathryn C. 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, speed 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--what 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 Honikman [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.

Atatlahuca 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 try to remember each time he comes to individual sounds such as *n, t, k, ɨ, u, a* 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 e* or *este este* in contrast to English speakers' *əəə*. As to rhythm, both Atatlahuca

Mixtec and Spanish exhibit syllable timing rather than stress timing as do English and Southern Tepehuán.

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 Chinantec, Yatzachi Zapotec, Atatlahuca 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 Tepehua where people generally do not speak softly.

The Mazatecs speak quietly. Women never raise their voices. In Huautla there is a large market, full of hundreds of people, but you cannot hear it until you are 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 Chinantec, Chiquihuitlán Mazatec, Coatzospan Mixtec, Quiotepec 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 Pike 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 phonostylistic 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, insulting, or impolite when speaking.

John Crawford reports that when he lived among the Mixe people, he could always tell when a visitor was leading up to asking to borrow money, as the visitor always used creaky voice. A mad, excited Mixe speaker used a monotone with a dive down at the end. For emphasis or excitement, the speech was breathy.

In Huautla 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 my:::: tur:::n no:::w." Urgency, on the other hand, is expressed by breathiness, as when impatiently calling for someone: "Vic̤to̤ria, Vic̤to̤riahhh!" 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 Díaz Mazatec, Alacatlatzala Mixtec, and Ocotlán Zapotec. The high pitch is sustained in Highland Totonac, without lowering. In Cuicatec and Cora (Mexico) and in Tucano (Colombia) 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 Ixtlán Zapotec, Northern Tlaxiaco Mixtec, and Atatlahuca Mixtec. In each of these the speech is rapid, and with narrowed pitch range. In Western Ixtlán 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 Chinantec the speech is stacatto. In Chiquihuitlán Mazatec there is exaggerated aspiration. Loudness, protracted syllables and some breathiness are reported for Ozumacín Chinantec. In Náhuatl of Tetelcingo and of Orizaba there is an abrupt cutoff of phrases and sentences preceded by abrupt downturn of intonation. Michoacan Náhuatl and Southern Tepehuán speakers talk quietly to their children. Tepetotutla Chinantec 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 Lacandón and Guelavía Zapotec.

Quite a few languages exhibit general fronting, or specific consonant changes such as palatalization. In Trique not only is there replacement of alveopalatals by fronted alveopalatals or dentals, but sometimes replacement of dentals by alveopalatals or fronted alveopalatals. Atatlahuca Mixtec *tʃ* is substituted for *ʃ*, *i* for *ɨ* and initial *s* of consonant clusters is dropped. In Coatzospan Mixtec *ð* becomes *l* and *ts* becomes *t*. In Veracruz Tepehua the consonant changes are: *ʃ* > *ṣ* *tʃ* > *ṭṣ* *ts* > *tʃ̣* *q* > *ḳ*. In Seri *s* > *ʃ̣*

3.2.3. Showing Sympathy

We have mentioned that lip rounding is used in Mazatec and Mixtec to show sympathy. In San Felipe Otomí and in Veracruz Tepehua it is used both to express and elicit sympathy.

Creaky voice is reported for Alacatlatzala Mixtec and Trique. In Trique falling pitch is superimposed on the tone system, and increased creaky voice occurs as the pitch falls; also the particle at the end of the sentence is lengthened. Choapan Zapotec and Highland Oaxaca Chontal are soft spoken. Lacandón 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 comadre or compadre. 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 offical or a witch doctor. The falsetto shows submissive attitude and sometimes fear. San Felipe Otomí 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 Díaz Mazatec, a language with all open syllables. In Alacatlatzala 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 Ozumacín Chinantec the lower pitch is accompanied by lower volume. Tlapanec 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. Chiquihuitlán 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 Otomí when pleading for mercy. For example, a young boy being scolded and threatened with a whipping might switch into falsetto.

3.2.7. Emphasis

Heavier 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 Díaz Mazatec. In Southern Tepehuán high pitch and lengthened vowels are used.

REFERENCES

[1] HONIKMAN, B. (1964). "Articulatory Settings", in *In Honour of Daniel Jones*", Abercrombie, David et al. London: Longmans 73-84.

[2] LAVER, J. (1980). "*The Phonetic Description of Voice Quality*". Cambridge: The University Press.

[3] PIKE, K. L. (1957). "Abdominal Pulse Types in Some Peruvian Languages", *Language*, 33.1.30-35.

[4] PIKE, K. L. and PIKE, E. G. (1983). "*Text And Tagmeme*", Norwood, New Jersey: Ablex Publishing Corporation. 76, 80-85.

RIEZ-VOUS EN HI! 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 greatest agreement on lexicalized laughter than on its real perception.

1. DU RIRE SÉMIOTIQUE AU RIRE SÉMANTISÉ

Le rire, phénomène paralinguistique (Crystal, 1969) [1] doit être envisagé comme une manifestation émotive. On peut alors l'analyser selon le modèle élaboré pour l'étude des émotions (Léon, 1976) [4] considérant d'une part un rire *brut index sémiotique de l'émotion et de la personnalité,* et d'autre part, un rire *socialisé signalant* une attitude.

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, coléreux, 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 [ksss]! (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 *amusés*.
- 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 *gêné et sexy* à 40%; un *bête 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	+	—
mélodicité	—	+
vocalité	—	+

On va ainsi du désordre à l'ordre. Les pulsions dont le rire est fait sont toujours présentes mais elles tendent à l'irrégularité 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 variantes; 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 F_0 oscille entre 100 et 168 Hz, $\mu = 112$, $\sigma = 29$ Hz. L'intensité générale est assez forte (32dB mais les pulsions, très vocaliques, ne sont séparées que par de faibles changement d'intensité ($\mu = 32.8$ dB; $\sigma = 2.2$ dB).
- *surpris*: se manifeste ici par un souffle suivi d'une partie sonorisée, légèrement nasale et de *montée* mélodique rapide (fig.2), caractéristique du patron prosodique de la surprise.

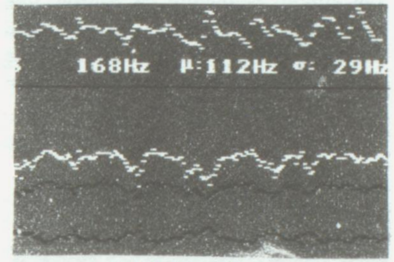


Figure 1: Rire "joyeux".

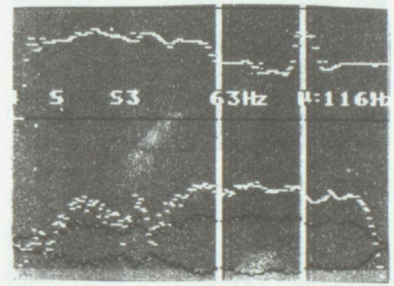


Figure 2: Rire "surpris".

- *ironique*: Comme pour la voix ironique, le rire ici montre des montées mélodiques accompagnées de chutes d'intensité ou d'une absence d'accroissement. La figure 3 indique un changement de +86 Hz pour une intensité décroissante de —1 dB.

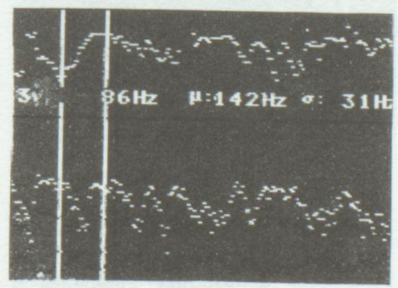


Figure 3: Rire "ironique".

- *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 sautes d'intensité importantes (μ = 27.1 dB, σ = 8.4 dB).

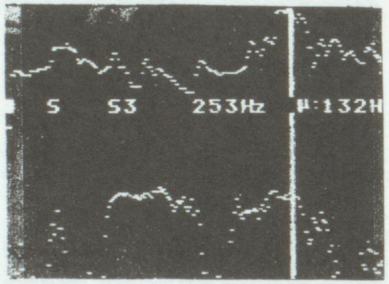


Figure 4: Rire "gêné et sexy".

- *bête et sadique*: Le patron rythmique commence par des pulsions longues (entre 20 et 25 cs) et se termine par une série de plus petites (8 à 10 cs). Le timbre est en [ɸ] et la mélodie plate (μ = 122 Hz, σ = 8 Hz). (Fig 5)

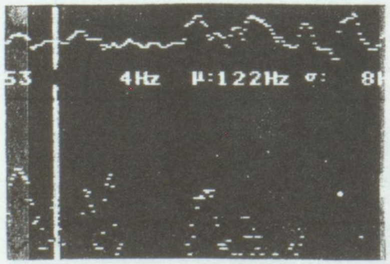


Figure 5: Rire "bête et sadique".

On a pu constater ici, dans les quelques patrons analysés, des configurations acoustiques analogues à celles relevées pour les *émotions* dans la parole (Fónagy, 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 hein! hé hé hé! hou hou hou!* 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% *joyeux* (4+4) 40%; *sarcastique* (2+2) 20%;

Ha ha ha!: garçon (6+10) 80%; *joyeux* (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%

Hé 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 hi*: *nerveux* (15%); *ironique* (10%).

- On voit très bien se dessiner dans l'imaginaire des sujets parlants le rire en *Hi 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 hein!* est jugé *réprobateur* (20%) ou *sarcastique* (80%) et également *supérieur*. Tous ces sèmes se rejoignent et confirment l'observation freudienne d'Ivan Fónagy [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%). Ici encore le *trait acoustique +aigu* du [e] a joué comme pour le [ɪ]. 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 recéler 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 la 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 sa fonction 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

[1] CRYSTAL, D. (1969), *Prosodic System and Intonation in English*, Cambridge U.K., University Press.

[2] FÓNAGY, I. (1983) *La vive voix*, Paris, Payot.

[3] LÉON, P. (1971), *Essais de phonostylistique*", Studia Phonetica, 4, Montréal, Paris, Bruxelles, Didier.

[4] LÉON, P. (1976), *De l'analyse psychologique à la catégorisation auditive et acoustique des émotions dans la parole*, *Journal de psychologie*, 3-4, 305-324.

[5] LÉON, P., R. DAVIS, et D. HEAP (1991), *Sémiotique et sémantique du rire* in *Information/Communication*, 12, numéro spécial: Sémio-sémantique et sémio-linguistique.

VARIABLES INTONATIVES CHEZ LA FEMME VENEZUELIENNE

Elsa Mora

Universidad de los Andes, Mérida, Venezuela

ABSTRACT

This paper analyses intonation in the registers of Venezuelan Women of different social class. Graphic representation of the results is done using the intonation model proposed by Fant (1984). Results show social class phonostylistics differences.

1. INTRODUCCION

- Les modes de créativité du langage sont d´une telle 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´énoncé 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 qui va au-delà du message réferentiel. "Bien souvent pourtant la fonction référentielle a une importance minime et le véritable message ne peut étre décodé que dans la parole proférée" (león 1979:159).

- Nous essaierons de montrer, dans cette étude, la fonction phonostylisque de certaines expressiones 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 sélectionné un ensemble d´énoncés qui "sembleraient" 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 intonatives initiales, médiales et finales, mais la courbe mélodique se trouve simplifiée par le fait qu´une séquence de syllabes atones, quel que soit leur nombre, est toujours représentée comme ayant la même quantité qu´une syllabe tonique. Ce type de représentation omet plusieurs phénomènes tels que la pente globale de la courbe mélodique à mesure que l´énoncé se déroule.
- Le systéme de notation de Fant permet de distinguer quatre niveaux significatifs de la courbe mélodique: un niveau bas (B) un niveau moyen (M) un niveau haut (H) et un niveau haut extrëme (H+).

4. ANALYSE DE L´ECHANTILLON

- De tout le corpus analysé, nous avons isolé, seulement à titre d´exemple, trois expressions non marquées par **le fait** phonostylistique que lón veut faire remarquer, et trois expressions marquées.

1.1. Graphiques de l´intonation "non marquée":

- Les figures (1), (2), (3), présentent l´intonation non marquée.

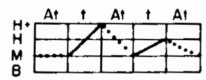


Fig. 1

TE INVITO SIN DUDA

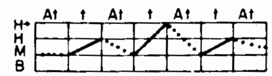


Fig. 2

LOS POBRES PAGAN CARCEL

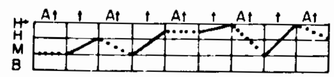


Fig. 3

HABLO CON TODA LA VERACIDAD POSIBLE

- Le graphique de la figure (1) présente une élévation initiale du ton partant d´un niveau moyen et arrivant au point le niveau élevé correspondant à la fin de la dernière tonique de l´énoncé pour terminer au niveau moyen en fin d´énoncé. Le point le plus élevé (H+) de ce grahique tonal ne coincide pas toujours avec la première tonique de l´énoncé, ainsi quón peut le remarquer sur les figures (2) et (3), il semblerait que le point le plus élevé va retomber sur la tonique sur laquelle on veut insister.

4.2. Graphiques de l´intonation "marquée":

- Les figures (4), (5) et (6) représentent l´intonation "marquée".

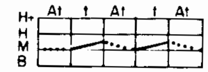


Fig. 4

YO SE LO DIJE

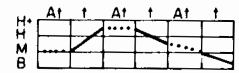


Fig. 5

PREFIERO VIAJAR EN AVION

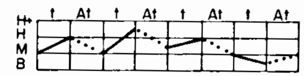


Fig. 6

SOLO USO PERFUMES FRANCESES IMPORTADOS

-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. La dernièr syllabe tonique peut varier d´une montée très faible (fig.4) à une descente considérable (fig 6).
- Sur les graphiques (4), (5),et (6), on peut remarquer un mouvement ascendant-descendant sensible, dû 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 sur nos graphiques, il nous faut singnaler un allongement sensible de la voyelle en syllabe accentuée, qui s´accompagne, dans la plupart 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 perçues comme l´emphase, les courbes descendantes, comme de la persuasion. La diversité de l´effet de l´intonation est telle que l´aspect sémantique de l´expression est plus important que celui des mots textuellement représentés.
- Les courbes intonatives marquées et non marquées sont des fragments de la parole qui tentent de compléter une expression dans une séquence organisée du discours global.

5. VARIABLES INTONATIVES ET FACTEURS PHONOSTYLISTIQUES.

- Des registres correspondant aux 30 locuteurs qui constituent la totalité de l´échantillon de ce travail, on a pris 1150 phrases déclaratives de manière à quantifier le pourcentage d´occurence de la courbe intonative marquée et non marquée.
- Le tableau 1 indique les résulats de cette analyse:

tableau 1

CI M	316	27,4 %
CI non M	834	72,5 %

- les données démontrent que dans le corpus étudié l´expression intonative marquées est beaucoup moins fréquente que la non marqué.
- Le tableau 2 indique le nombre et le pourcentage des occurences eb fonction de la classe sociale:

tableau 2

	F	%	M	%	D	%
CI M	130	(41)	72	(23)	114	(36)
CI non M	265	(32)	454	(54)	115	(14)

- Comme on le remarquera sur le tableau 2, le nombre d´occurences de la variable marquées décroît dans l´ordre suivant: classe favorisée, défavorisée, moyenne.
- Il faut maintenant insister sur le fait que la structure sociale de cette communauté linguistique est déterminée, dans son niveau élevé, par une position de pouvoir social qui n´est pas indicateur de niveau cultural. La classe moyenne est constituées, en majorité, par des femmes de formation universitaire intégrée au marché du travail.
- L´utilisation de la variable marquée est une manifestation - tout à fait consciente d´une fonction qui nést pas référentielle mais expressive; le sujet toujours à l´affût d´une réaction favorable de l´interlocuteur, cherche en quelque sorte à éluder ou à manipuler par séduction, c´est pourquoi elle se présente, dans la plupart des cas, comme une expression bien délimitée dans le discours, soigneusement articulée et prononcée d´une voix douce.

6. CONCLUSIONS

1- L´intonation marquée phonostylistiquement se différencie de celle que l´on considère non marquée par des oppositions de hauteurs, de quantité (bien qu´on ne puisse pas le noter sur nos graphiques) et de variabilité de l´inclinaison en fin d´énoncé.

2. Nous avons pu noter la corrélation entre la courbe intonative marquée et les traits phonostylistiques tant dans des fragments de discours cohérents et "cohesionnés" que dans ceux de la langue courante. Ce Type d´intonation fonctionne comme une structure d´utilisation consciente, bien découpée et délimitée dans le discours et possède une finalité spécifique.

3. Le facteur social est pertinent dans l´emploi plus ou moins important des variables intonatives. La variable marquée est employée principalement par la classe favorisée et, dans une moindre mesure, par la classe moyenne. La structure sociale de cette communaut´linguistique indique que le niveau élevé correspond à un niveau économique et non pas intellectuel. Le niveau moyen corespond au groupe intellectuel dans son ensemble. La classe défavorisé est celle qui accés ni au puovoir économique ni au pouvoir intellectuel. Cettestructuration sociale permet que l´emploi de certains patrons intonatifs, comme celui que nous avons appelé "intonation marquée" serve à éluder ou à manipuler une situation gräce à la séduction implicite dans la mélodie que ce type d´intonation renferme.

4. On a pur remarquer que l´emploi de la variable intonative marquée ne correspond pas exclusivement aux femmes d´un niveau déterminé dans la société mais que certains groupes peuvent se différencier par l´emploi plus ou moins prononcé de cette variable.

- Ce n´est donc pas l´objectivité de la fonction référentielle qui donne toute l´information, c´est le message phonostylistique qui rend compte du sens occulte de l´expression et qui nous permet de conclure que toute parole est revëtue d´intention.

7. REFERENCES

|1| CHELA-FLORES, B. (1990),"*Estudio socio-linguistico de la entonacion en el habla de Maracaibo*". Trabajo presentado en el XI ENDIL San Cristobal.

|2| FANT, L. (1984), "*Estructura informativa en español: estudio sintactico y entonativo*", *Acta Universitatis Upsaliensis.*

|3| LEON, P. (1971), "*Essais de phonostylistique*". Studiä Phonetica, 4, Montréal, Paris, Bruxelles Didier.

|4| LEON, P., FAURE,G., RIGAULT, A. (edit), (1970),"*Prosodic feature analysis. Analyse des faites prosodiques*", Studia Phonetica 3, Montreal, Paris, Bruxelles,Didier.

|5| LEON, P. et M. Rossi (édit), (1979), "*Problèmes de prosodie.Vol II: Experimentations, modèles et fonctiones*", Studia Phonetica 18, Montreal, Paris, Bruxelles,Didier.

|6| MORA, E. (1990), "*Phonostylistique de l'intonation: differenciation dues au milieu social et au sexe del locuteurs*", Revue québecquoise de lingusitique. Vol.19, nº 2, 73-92.

ETUDE DES PARAMETRES TEMPORELS DES VOIX SANS LARYNX

L. Crevier-Buchman

U.F.R. Linguistique, Université Paris VII, France.
Service O.R.L. Phoniatrie, Hôpital Laënnec, Paris.
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-esophagial voices using plumonary air were able to preserve the rhythm and the syntactico semantic structure of their speeches, whereas speakers with esophageal voices often needed to insuflate the esophagus 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 (PP) (Groupe III). Deux signaux ont été enregistrés : le signal acoustique et le signal électroglottographique (EGG). 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 aucun regret".
- le calcul du nombre de syllabes lu par minute lors de la lecture d'un texte "Grand-mère raconte" (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 les éléments sonores 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 lue 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 Edit" et "Signalize". Les données statistiques ont été analysées par le programme PCSM 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 existe une 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 tenue 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 (Groupes II et III), avaient des durées d'émission vocalique allant de 5 à 11 sec. La différence de TMP n'était pas significative entre les Groupes II et III (p=0.45). 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 proche de la normale.
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=0.004). En effet, lorsque l'on compare les 3 Groupes de laryngectomisés entre eux, ils avaient une durée de phrase équivalente quelque soit leur mode de production.
3.2.2. Durée de phonation
Pour les patients laryngectomisés elle était proche de celle des témoins. Il n'y avait pas de différence significative globale.
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=0.004).
- 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=0.42), alors qu'elle était significative avec le Groupe III.
- Comparaison avec les témoins : il n'existait pas de différence significative entre le Groupe III et le Groupe témoin (p=0.2), alors que la durée des pauses était toujours supérieure à la normale pour les Groupes I et II.
3.3. Etude longitudinale de patients
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 les constatations suivantes : diminution de la durée totale de la "phrase", en relation exclusivement avec un raccourcissement de la durée de la phonation ; en effet, les temps de pause nécessaires aux reprises inspiratoires et aux inructations étaient peu compressibles. L'autre patient avec FTO ne parlait qu'en voix chuchotée lors du 1er 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.
3.4. Le débit phonatoire
Il a été calculé à partir du nombre de syllabes lues par minute. On a pu constater une réduction du nombre de syllabes lues par minute, chez tous les 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 des Groupes I et II.

4. DISCUSSION

4.1. Le temps maximum de phonation

Le temps maximum de phonation reflète les capacités physiologiques d'émission prolongée de voisements. 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 autonomie expiratoire 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 resistance 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 néoglotte et l'adaptation de la pression sous néoglottique, responsable de ces variations temporelles.

4.2. En situation de parole

Notre étude a mis en évidence que les patients laryngectomisés élaborent une stratégie de lecture qui se ferait 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 inructations 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 auraient tendance à dire 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 maniérismes et les particularités du locuteur (2).

Tableau des résultats

m : moyenne ; DS : Déviation Standard (écart type) ; p : probabilité
V.O. : Voix Oesophagienne ; P.P. : Prothèse Phonatoire
F.T.O. : Shunt Trachéo-Oesophagien autocontinent

		V.O.. Groupe I	F.T.O. Groupe II	P.P. Groupe III	Témoins	p
Temps maximum de Phonation (secondes)	m	2.00	7.66	8.25	11.28	0.023
	DS	0	225.09	262.99	242.9	
Durée de la Phrase (secondes)	m	7.71	7.44	5.9	4.97	0.004
	DS	9.89	146.68	68.49	61.37	
Durée du Voisement (secondes)	m	4.95	5.05	5.02	4.33	0.353
	DS	91.92	119.21	63.39	51.47	
Durée des Pauses (secondes)	m	2.75	2.30	0.87	0.64	0.004
	DS	83.43	79.29	36.17	33.5	
Nombre de syllabes/minute	m	164	141	167	223	0.006
	DS	52.32	28.68	31.55	16.46	

6. REFERENCES

(1) BRASNU, D., STROME, M., CREVIER-BUCHMAN, L., PFAUWADEL, M.C., LACCOURREYE, H. (1989), "Voice evaluation in myomucosal shunt after total laryngectomy ; comparison with Esophageal speech" Am. J. Otolaryngol., 10, 267-272.
(2) GUNN, D.A., MONTAGNE, J.C., TORGERSON, J.K. (1979), "A comparison between laryngectomized and non-laryngectomize male esophageal speakers on selected auditory perceptual parameters of esophageal speech", Folia Phoniat., 31, 167-176.
(3) NIEBOER, G.L., DE GRAAF, T., SCHUTTE, H.K. (1988), "Esophageal voice quality judgements by means of the semantic differential" Journal of Phonetics, 16, 417-436.
(4) PFAUWADEL, M.C., CREVIER-BUCHMAN, L., BRASNU, D. (1991), "Speech versus voice. A more pertinent approach to evaluate alaryngeal speakers", Ear Nose Throat J. (in press).
(5) PINDZOLA, R.H., CAIN, B.H. (1989), "Duration and frequency characteristics of tracheoesophageal speech", Ann. Otol. Rhinol. Laryngol., 98, 960-964.
(6) SEDORY, S.E., HAMLET, S.L., CONNOR, N.P. (1989), "Comparison of perceptual and acoustic characteristics of tracheoesophageal and excellent esophageal speech", Journal of Speech and Hearing Disorders, 54, 209-214.

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 trying to establish the acoustic cues for naturalness in laryngectomee speech and what constitutes the typical alaryngeal voice quality. Other tasks include intelligibility and acceptability ratings with professional 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, either by air that is insufflated 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 (48 - 80 years) have been analysed. 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 /papapa:/, embedded in a carrier phrase. Subjects were asked to produce these words at three loudness levels, subjectively 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 H_2O, for the TE-speakers 22 cm H_2O and for the normal speakers 7 cm H_2O. This result compares favourably with what is known from investigations of sound 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.

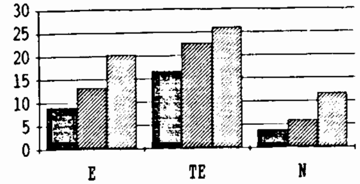


Figure 1. Pressure values in cm H_2O during the production of /p/ for E-, TE- and normal laryngeal speakers, at three loudness levels, weak, normal and strong. (3 subjects in each speaker group; 15 samples/displayed value)

4. INVERSE FILTERING AND VOICE QUALITY
By means of inverse-filtering of the air flow during phonation, the aperiodicity of the wave shapes was analysed and correlated to perceived voice quality.
Flow glottogram curves were obtained for many of the alaryngeal speakers, although they showed a great deal of irregularity. In Figure 2, two examples of automatic inverse filter analysis are shown [2].

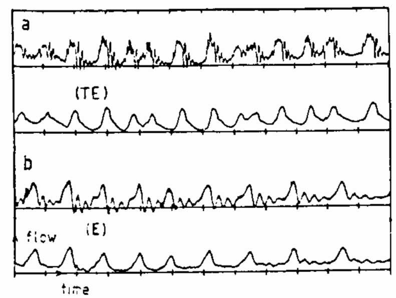


Figure 2. Flow registrations (upper curve) and corresponding inverse filtered flow glottogram (lower curve) of vowel pulses in /papapa:/, uttered by a TE-speaker (a) and an E-speaker (b).

Inspection of the unfiltered speech wave 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.

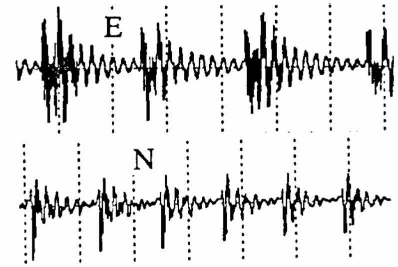


Figure 3. Speech wave oscillograms of vowel samples for an E-speaker and a normal speaker (N).

5. LONG TIME AVERAGE SPECTRA
Long-time-average spectra of these voices have been derived and analysed. A reading of text passage of approximately 45 secs was used as analysis material. The signal was fed into a Hewlett Packard 3562A 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 measure of 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.

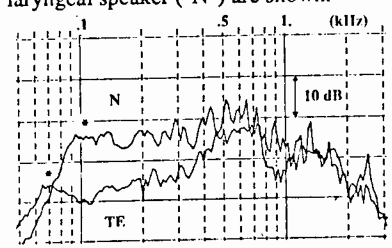


*Figure 4. Long Time Average spectra of a read text passage by a TE-speaker and an N-speaker. The level of the fundamental, L0, is indicated by *.*

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 speakers, suggest that the L1-L0 difference is larger for the alaryngeal voices than for normal voices, i.e. the level of the fundamental is very weak in the alaryngeal voices, see Figure 5. Moreover, it does not vary with loudness to the same extent as in normal laryngeal voices [11].

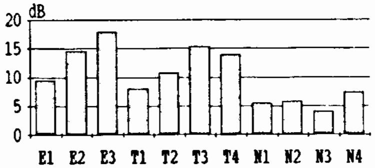


Figure 5. Difference data of total level (L_{tot}) to the level of the fundamental (L0) derived from LTAS-analysis of a text passage, read by three E-speakers, four TE-speakers and four N-speakers.

6. PROSODY

Pitch and duration cues

Prosodic studies of intonation patterns and word emphasis related to overall 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 alaryngeal speakers, one female E-speaker and one male TE-speaker. As can be seen the pitch patterns are varying in much the same fashion as for normal laryngeal phonation. Note the high pitch produced by the female E-speaker and the very low pitch produced by the male TE-speaker.

Using automatic pitch extraction algorithms, these voices are difficult to analyse, depending on the low voice registers, and irregular vibration patterns. The analysis was made by trying different pitch extraction algorithms, developed by Liljencrants [8]. Visual inspection was also performed on spectrograms and oscillograms. For some of the aperiodic voices it was very difficult to identify any periodic component, although it was still perceivable.

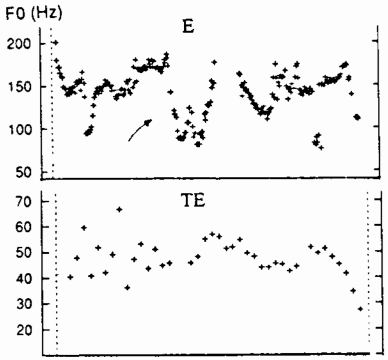


Figure 6. Pitch analysis of a sentence with emphatic word accent (female E-speaker) and a question (male TE-speaker).

The voice quality of the female E-speaker above was quite strained and rough, although she managed well to produce acceptable pitch patterns and had a quite high pitched voice (mode value 148 Hz). The voice breaks easily into a much lower register. Reasons for this are probably to be found in the changes of pressure conditions at the voice source, caused by consonantal constrictions. In Figure 6, such a break occurs at the arrow, /o:/ followed by /l/. The "diplophonic" rough character that this voice exhibits has also been noticed in a male esophageal speaker with a high pitched voice (mode value 121 Hz). It is as if the vibrating PE-segment is very sensitive to the right level of driving pressure; a too weak pressure will not start any oscillatory process and too much force may create deviant vibratory frequencies.

Typically, a very low pitch is common. The problem often is that the aperiodicity creates noise, overlaid on the fundamental. See the second pitch curve, displayed in Figure 6. Although varying in a normal fashion, 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 alaryngeal 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 alaryngeal groups are evident. Firstly, the alaryngeal 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, laryngeal voices will have a more pronounced fundamental if they phonate at low intensities. The same does not happen for the alaryngeal speakers.

8. CONCLUSIONS

It was found that the alaryngeal voices, E-voices and TE-voices were characterized by a weak level of the fundamental compared to normal laryngeal voices. Other, more detailed voice source characteristics, such as inverse filtered flow registrations displayed strong irregularities for the alaryngeal voices.

Acknowledgement: This work was financed by research grants from the Swedish Cancer Society and the Swedish Council for Social Sciences.

REFERENCES

[1] Askenfelt A., Hammarberg B. (1986): "Speech waveform perturbation: a comparison of seven measures", *J of Speech & Hearing Research 29*, pp. 50-64.

[2] Gauffin J., Hammarberg B., Imaizumi, S. (1986): "A microcomputer based system for acoustic analysis of voice characteristics", *Proc ICASSP Tokyo Vol 1*, pp. 681-684.

[3] Hammarberg, B. (1986): "Perceptual and acoustic analysis of dysphonia". *Studies in Logopedics and Phoniatrics No 1, Huddinge Univ Hospital.* Doct Dissertation.

[4] Hammarberg, B., Lundström E., Nord L. (1989): "Aspects of laryngectomee speech communication - a progress report." *STL-QPSR 1*, pp. 175-178.

[5] Hammarberg, B., Lundström, E., Nord, L. (1990): "Consonant intelligibility in esophageal and trachoesophageal speech. A progress report", *Phoniatric & Logopedic Progress Report 7*, Huddinge Univ Hospital, pp. 49-57.

[6] Hammarberg, B., Lundström, E., Nord, L. (1990): "Intelligibility and acceptability of laryngectomee speech", *PHONUM, Reports from the Department of Phonetics, Univ of Umeå, 1*, pp. 88-91.

[7] Hammarberg, B., Nord, L. (1989): "Tracheo-Esophageal Speech, Esophageal Speech and Artificial Larynx Speech - Acoustic and Perceptual Aspects", *Proc XXI Congr Int Assoc Logopedics & Phoniatrics, Prag August 6-10*, pp. 426-428.

[8] Liljencrants, J. (1991): "Manual for speech analysis software on the Apollo computer" (unpubl manuscript). Dept Speech Comm & Music Acoustics.

[9] Liljencrants, J. (1991): personal communication

[10] Nord L., Hammarberg B. (1988): "Communicative aspects of laryngectomee speech. Some preliminary phonetic-acoustic results from an interdisciplinary project", *STL-QPSR 4*, pp. 31-49.

[11] Nord L., Hammarberg B. (1989): Analysis of Laryngectomee Speech - A Progress Report. In *'Eurospeech 89', European Conf on Speech Comm & Technology (ed. by J.P. Tubach & J.J. Mariani), Vol II*, pp. 493-496, CPC Consultants Ltd, Edinburgh, UK.

[12] Robbins J., Fischer H., Blom, E., Singer, M. (1984): "A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production", *J of Speech and Hearing Disorders 49*, pp. 202-210.

[13] Rothenberg, M. (1973): "A new inverse-filtering technique for deriving the glottal air flow during voicing", *J of Acoust Soc of America 53*, pp. 1632-1645.

ON USING INTENSITY AS A CODING PARAMETER IN TACTILE SPEECH STIMULI: PSYCHOPHYSIOLOGICAL DISCRIMINABILITY EFFECTS

H. G. Piroth and S. Gfroerer

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 psychophysiological effects that show to be involved in intensity perception, namely the order effect and masking.

1. INTRODUCTION

Not only in psychophysiology, but also in application-oriented research to establish electrotactile 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 [13] we assume that a suprasegmental component (stress, intonation) could be added to the feature coding method by superimposing intensity variations on the segmental stimuli [9].

Classical investigations on electrotactile intensity perception discuss the number of possible steps that can be discriminated between absolute threshold and pain. Lindner 1937 [5] has reported that the pain threshold is reached at approximately four times absolute threshold. At a frequency of 400 Hz he situates absolute threshold at about 0.8 mA, pain threshold at 4.7 mA with 27 discriminable steps in between. Schöbel 1936 [11] determined a difference limen of 4 to 5% in normal hearing subjects, Anderson and Munson 1951 [1] of 2 to 5% in the frequency range between 100 and 5000 Hz. Hawkes 1959 [4] measured a limen of 5.3% at an intensity of 120% and of 3.8% at 200% above absolute threshold.

A pilot experiment with more complex stimuli [10] using electrocutaneous pulse train sequences 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 Electrocutaneous Stimulation SEHR-2. Four rows of electrode pairs were fixed along the dorsal, ulnar, volar, and radial sides of the Ss' left forearm. (See [13] for details and illustrations.)

3. STIMULI

Four complex stimuli were constructed with pulse train sequences as their basic part consisting of three bipolar pulses with a rectangular part in one and hyperbola-shaped part in the other polarity, resulting in a d.c.- component equaling 0. The pulse repetition rate was 400 Hz. In stimulus I 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 of the arm and proceeded to the dorsal side. Then, without an additional pause a longitudinal sequence of pulse trains was presented oscillating between the distal electrode pair on the dorsal side and the neighbouring dorsal electrode pair fixed 4 cm apart in proximal direction. This sequence started at the more proximal place and consisted of eight pulse trains separated by an interval of 20 ms after each pulse train.

In stimulus II the order of the two parts was changed, thus it started with the longitudinal part and ended with the surrounding one. For stimulus III the complete surrounding part of the pattern was presented to the ring of electrode pairs placed 4 cm apart from the distal ring in proximal direction. The longitudinal part that followed remained the same as in stimulus II, but started from the distal electrode pair.

In stimulus IV again, the order of the two parts of stimulus III was altered.

According to the feature coding method discribed in [13] stimuli I to IV are the tactile equivalents of /fi:/, /i:f/, /ʃi:/, and /i:ʃ/. To determine stimulus intensities each S underwent a calibration procedure before each test. The Ss had to adjust absolute threshold and the threshold of annoyance four times for each place of stimulation in a mixed ascending and descending procedure using the basic parts of the stimuli presented repeatedly and separated by an interval of 50 ms. Nine intermediate intensity values were calculated corresponding to the absolute threshold +10%..., +90% of the difference between both thresholds. Accordingly, seven versions of the five stimuli were defined with the intensity of the rectangular parts of the pulses in the vowel pattern (i.e. the longitudinally moving part) set to the 3rd to 9th intensity value as calculated. The intensity of the consonantal pattern (the surrounding one) was two steps (20% of the threshold difference) lower than that of the vocalic part.

4. PROCEDURE AND SUBJECTS

Stimuli were arranged in pairs to yield a two-step discrimination test for stimulus intensities. All pairs contained two repetitions of the same stimulus with an interval of 1 s within the pair. Five pairs were built for each stimulus with higher intensities in the second stimulus (intensity values 3-5, 4-6, 5-7, 6-8, 7-9) and the five corresponding pairs with lower intensities in the second stimulus.

In this way a 4x2x5-factorial test design was constructed with 4 stimuli, 2 orderings (ascending and descending intensities) and 5 intensity levels.

One subtest included 10 repetitions of pairs of stimuli I and II (/fi:- fi:/ and /i:f-i:f/) in randomized order, the other subtest of stimuli III and IV (/ʃi:-ʃi:/ and /i:ʃ-i:ʃ/), resulting in 200 pairs for each subtest. 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 intensitive stimulus of each pair on an answer sheet.

5. RESULTS

Tab. 1 gives the results of a 4x2x5-factorial MANOVA (SPSS; [6] 1975) with stimulus, ordering and intensity level as factors. The overall discriminability was 80.53% showing that the intensity differences were well-recognizable. The MANOVA calculation yielded a significant stimulus effect ($p<0.05$) and a highly significant interaction of intensity level and ordering ($p<0.001$). It can be seen from Fig. 1 that dicriminability increases with intensity level for the series of 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 a posteriori test showed significant differences ($p<0.05$) between /ʃi:/ and /fi:/, as well as between /ʃi:/ and /i:ʃ/, and /ʃi:/ and /i:f/. /i:f/ and /i:ʃ/ showed a slight ($p<0.10$) tendency effect (Tab. 2).

Table 1
Results of the Statistical Analysis

Factor	d.f.	F	p	
ITEM	19,1	7.63	=0.04	*
ORDERING	19,1	0.09	=0.78	n.s.
LEVEL	19,4	2.44	=0.07	n.s.
IT x ORD	19,1	0.37	=0.59	n.s.
IT x LE	19,4	0.58	=0.67	n.s.
ORD x LE	19,4	12.42	<0.001	***
ITxORDxLE	19,4	1.00	0.42	n.s.

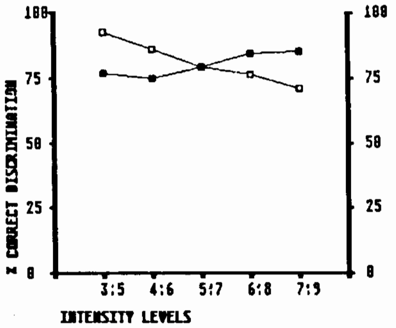


Figure 1: Discriminability dependent on intensity levels (full squares: ascending; open squares: descending)

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. [12]). 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 Piroth&Tillmann 1987 [8], thus it is clear now that duration as well as intensity perception of electrotactile 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 /ʃi:/.

The rank order of the stimuli shows that intensity discrimination tends to be better with /f/ than with /ʃ/ and better with VC than with CV. Similarly, Piroth 1986 [7] had shown that identification of tactile vowels is higher in VC and identification of consonants is higher in CV-syllables. Those effects

Table 2
Discrimination Dependent on Stimuli

	/i:f/	/fi:/	/i:ʃ/	/ʃi:/
%	83.88	81.49	81.00	75.75

/i:ʃ/ > /ʃi:/
/fi:/ > /ʃi:/
/i:f/ > /ʃi:/

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 in /ʃi:/ could then be explained if /ʃ/ had a stronger masking effect than /f/ in CV- stimuli. For such an explanation the central representations of the stimuli instead of their peripheral characteristics have to be taken into account. Within the frame of this experiment only a first speculative approach to such an explanation can be proposed: The basic units of the stimuli (pulse trains) were identical in all cases, but they differed in their temporal and spatial relations. Because of the somatotopic representation of body sites the spatial relations should be preserved in building the central representation. But since more distal and more proximal places were stimulated in the "consonantal" (circumferent) parts of the stimuli, the conduction velocities in the nerve fibres may become relevant to determine the central temporal relations.

In CV-stimuli the interval between the last pulse train of /f/ or /ʃ/ and the first of /i:/ is 15 ms. The distance between the corresponding places of stimulation is 4 cm, but in /fi:/ the place changes in proximal, in /ʃi:/ in distal direction when proceeding from the "consonantal" to the "vocalic" part. Relying on the values given in the literature ([2,3]) 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 stimuli is based on information processed via thin unmyelinated C- fibres with a conduction velocity of no more than 2.55 m/s the temporal intervals at the points of central occurance of two successice pulse trains at places 4 cm apart from one another differ from the peripheral interval by at least 15.7 ms. Thus, in /ʃi:/ with /ʃ/ being presented at more proximal places the inter-pulse-train interval is centrally doubled (15 ms + 15.7 ms = 30.7 ms), and in /fi:/ 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 conclude:

(i) /i:ʃ/ and /i:f/ cannot cause forward masking, since the vocalic part is presented first (81.00% and 83.88% correct discrimination).

(ii) /ʃi:/ produces a forward masking effect, since the central representation of /ʃ/ is built up before the representation of /i:/ is evoked (thus, only 75.75% correct answers).

(iii) For /fi:/, the representation of both parts are not separate, but as the central point of occurrance of the last pulse train of /f/ is nearly identical with that of the first in /i:/ 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 psychophysiological experiments are mandatory.

7. REFERENCES

[1] ANDERSON, A. & MUNSON, W. (1951), "Electrical excitation of nerves in the skin at audiofrequencies", *J.Acoust. Soc.Am. 23*, 155-159.

[2] BURGESS, P. & PERL, E.(1973), "Cutaneous mechanoreceptors and nociceptors", in: A. Iggo(ed.), *"Handbook of sensory perception II. Somatosensory system"*, 29-78.

[3] GASSER, H.(1955), "Properties of dorsal root unmedullated fibers on the two sides of the ganglion", *J.Gen.Physiol. 38*, 709-728.

[4] HAWKES, G.(1959), "Cutaneous discrimination of electrical intensity", *Phil. Diss.* Univ. of Virginia.

[5] LINDNER, R.(1937), "Physiologische Grundlagen zum elektrischen Sprachetasten und ihre Anwendung auf den Taubstummenunterricht", *Z.Sinnesphysiol. 67*, 114-144.

[6] NIE, N. et al.(1975), *"SPSS Statistical Package for the social sciences"*, New York.

[7] PIROTH, H.(1986), "Electrocutaneous syllable recognition using quasiarticulatory coding of stimulus patterns", *J.Acoust.Soc.Am. 79, Suppl.1*, S73.

[8] PIROTH, H. & TILLMANN, H. (1987), "An order effect in pulse train discrimination as a case of time order error", *Proc. 11th ICPhS, Vol. 5*, 50-53, Tallinn.

[9] PIROTH, H.(1989), "Tactile recoding of phonological features in a system for electrocutaneous substitution of speech for the deaf", *Magyar Fonetikai Füzetek 21*, 188-191.

[10] PIROTH, H.(1991), "Intensity discrimination and identification in electrotactile pulse train sequences", to appear in *Forschungsberichte des Instituts f. Phonetik u. Sprachl. Komm. Univ. München (FIPKM) 29*.

[11] SCHÖBEL, E.(1936), "Versuche über Intensitätsunterscheidung beim 'elektrischen Tasten' verschiedener Frequenzen", *Z.Sinnesphysiol. 66*, 262-273.

[12] STOTT, L.(1933), "The discrimination of short tonal durations", *Diss.* Univ. of Illinois.

[13] PIROTH, H. & TILLMANN, H. (1991), "Articulation-based tactile speech for the deaf: a complete set of tactile segmental features for German", *Proc. 12th ICPhS*.

SPEECH PERCEPTION ABILITIES OF PATIENTS USING COCHLEAR IMPLANTS, VIBROTACTILE AIDS AND HEARING AIDS

Eva Agelfors and Arne Risberg

Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Box 70014, S-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 listener 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 is more or less impossibly. 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 material does not exist in any language and the phonological characteristics from one language to another make the interlingual 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 tactual modality [6]. A few studies have also reported good tactual 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 (Vt:1-Vt:8). Three deafened adults (Vt:1-Vt:3) 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 vibrator in their left hand.

In the cochlear-implanted group, six subjects participated (Ci:1-Ci:6). Two subjects, Ci:1 and Ci:2, were implanted with a single-channel extra cochlear implant (Wien/3M) and four subjects were implanted with a multichannel intracochlear 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 (H1:1-H1:4) and (H2:1-H2:7). 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 aid. Although all subjects were profoundly impaired, the subjects were not equivalent audiometrically. For that reason they were divided into two groups: group H1 with mean hearing-loss at frequencies 500, 1000 and 2000 Hz of 104 dBm, 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 (Lp1-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 speech tracking. The segmental test used a set of 16 vCv 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 is a closed-set test battery, presented as a two alternative forced-choice task. The specific prosodic features tested were: number of syllables, vowel-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 verbatimly. 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 session to session. During each test session, tracking was performed for a total of ten minutes under each of two conditions: (a) lip-reading plus aid and (b) lip-reading 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 text material 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 situations. Each subject was tested individually. The test order was the same for all subjects: vCv-syllables, prosody and speech tracking. Each test started with the combined situation.

The normally hearing subjects (Vt-group) were masked by earplugs and pink noise in the test situation with lip-reading and aid. During the speech tracking situation they were sitting in a sound-attenuating test-room and viewed 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 matrixes were constructed for each individual and for each situation. 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 vCv-syllables displayed for each group of subjects in the two conditions are shown in figure 1.

aid". The recognition of distinctive features was improved with the aid especially by groups of subjects Lp-500, H2 and Ci. The subjects received very little information about voicing when they were only lip-reading, but the high proportion of information transferred about voicing in the combined situation shows that the devices provided strong cues of low-fundamental frequency for all subjects.

Results obtained from the suprasegmental test show that mean score of 78.2% correct, (sd. 5,6%), is greater than chance level (50%) for all groups in condition "vision only". In the condition "vision plus aid", the vibrotactile aid transmitted no added information to visual cues. Suprasegmental features were very well perceived by all hearing aid users and by normally hearing subjects.

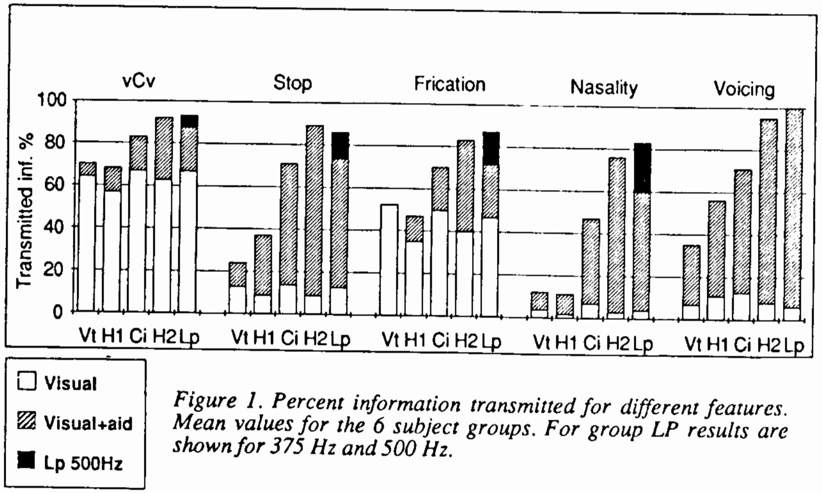


Figure 1. Percent information transmitted for different features. Mean values for the 6 subject groups. For group LP results are shown for 375 Hz and 500 Hz.

All subjects performed comparably in the "vision only" condition. There is no difference between normally hearing and hearing impaired subjects in this test condition. Two of the subjects, Ci:1 and Vt:1, are excellent lipreaders with more than 40 words/min in the speech tracking test. In spite of this, they did not achieve better results on the "visual only" consonant test. As expected, in "vision only" condition, the place feature was correct most often followed by frication. All groups of subjects, have got some improvement in the condition "visual plus

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 is 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.

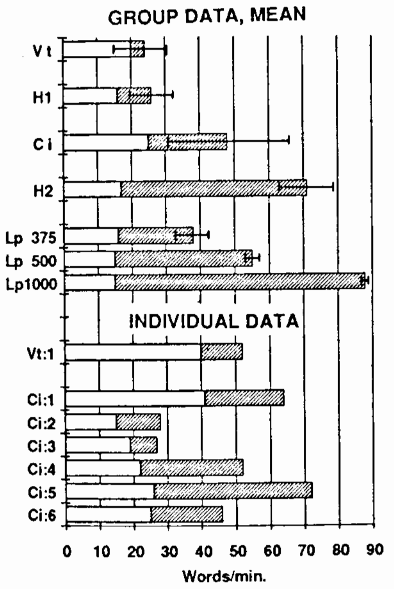


Figure 2. Results from speech tracking. Mean values for the different groups and individual values for the best tactile subject and the 6 cochlear implant subjects.

The individual data (Ci:1) in fig.2 shows that speechunderstanding performance for one single-channel cochlear implant subject can be as good as those obtained with multichannel cochlear implants. The difference between better and poorer patients is the detection/resolution of high-frequency components is reported by Dorman [2]. Responses of the subjects with hearing aids and with residual hearing (H2) were consistently superior to those subjects with implants or vibrotactile aids. The wide variation in responses of the cochlear implant users indicates the necessity of carefully evaluating each implant user.

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

[1] AGELFORS, E. & RISBERG, A. (1989) *"Speech feature perception by patients using a single-channel Vienna 3M extracochlear implant"*, STL/QPSR 1/89, 145-149.
[2] DORMAN, M., SOLI, S., DANKOWSKI, K., SMITH, L., (1990), *"Acoustic cues for consonant identification by patients who use the Ineraid cochlear implant"*, JASA, 88, 5, 2074-2079.
[3] DOWELL, R., MECKLENBURG, D., CLARK, G., (1986) *"Speech recognition for 40 patients receiving multichannel cochlear implants"*, Arc. of Otolaryngology, 86, 112, 1054-1059.
[4] de FILIPPO, C.L., SCOTT, B.L., (1978), *"A method for training and evaluating the reception of ongoing speech"*, JASA, 63, 1186-1192.
[5] MILLER, G.A., NICELY, P.E., (1955), *"An analysis of perceptual confusions among some English consonants"* JASA, 27, 338-3352.
[6] PLANT, G. (1986), *"A single-transducer vibrotactile aid to lipreading"*, STL/QPSR 1/86, 41-63.
[7] TYLER, R., MOORE, B.C.J., KUK, F., (1989) *"Performance of some of the better cochlear-implant patients"*, JSHR, 32, 887-911.
[8] WEISENBERGER, J.M., BROADSTONE, S.M., SAUNDERS, F.A., (1989) *"Evaluation of two multichannel tactile aids for the hearing impaired"*, JASA, 86, 1764-1775.

CHANGES IN SPEECH BREATHING FOLLOWING COCHLEAR IMPLANT IN POSTLINGUALLY DEAFENED ADULTS

Harlan Lane*, Joseph Perkell, Mario Svirsky, Jane Webster

Massachusetts Institute of Technology,
Cambridge, Massachusetts, U.S.A.

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-minutes 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 Ineraid 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 (Respitrace). 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 Respitrace 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

* Also of Northeastern University, Boston, Massachusetts

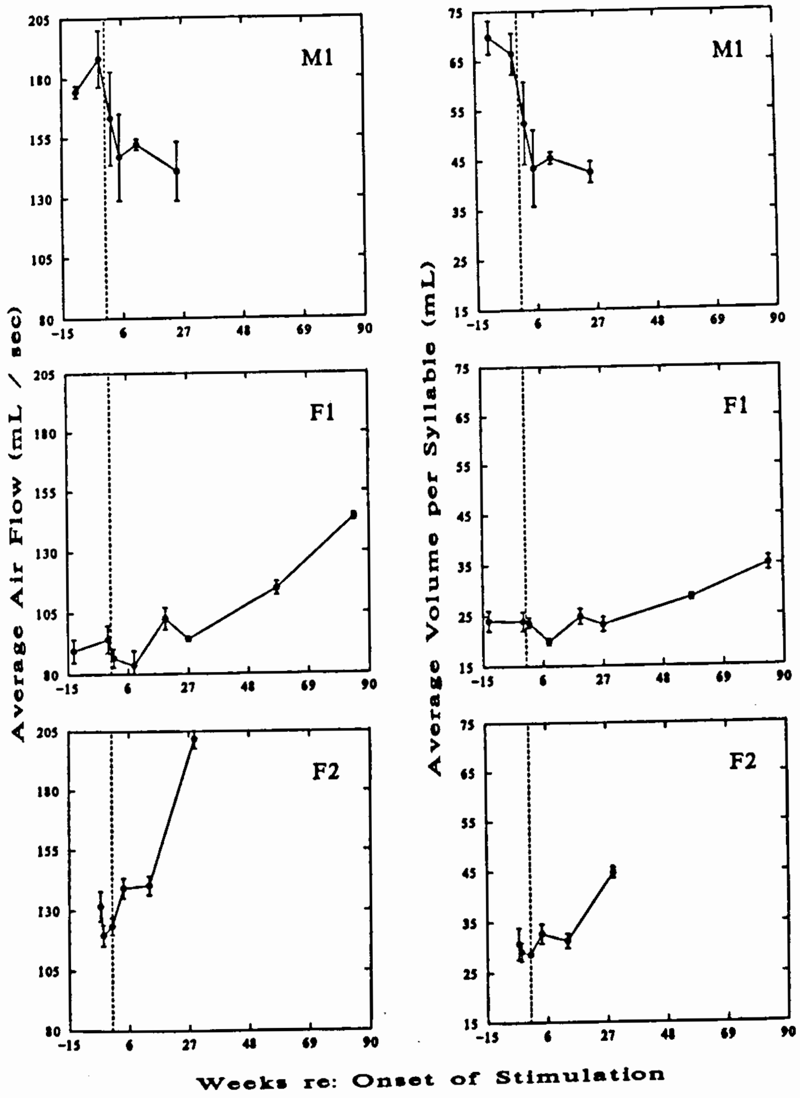


Figure 1. Average airflow (left column) and volume of air expended per syllable (right) measured while postlingually deafened adults read passages three times in each of two sessions prior to receiving stimulation from a cochlear prosthesis, and in four (M1,F2) or six sessions (F1) following onset of stimulation. (The vertical lines show when the processor was turned on.) Each point is the mean for three passages. Vertical bars show +/- one standard deviation of the passage means around the session mean.

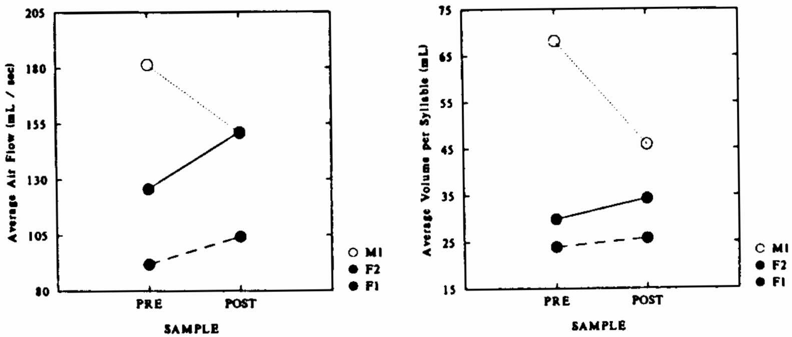


Figure 2. Average airflow and volume of air expended per syllable before and after cochlear prosthesis with postlingually deafened adults.

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. Insofar 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 of 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 normalcy. 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

[1] LANE, H., PERKELL, J., SVIRSKY, M. & WEBSTER, J. (in press). Changes in speech breathing following cochlear implant in postlingually deafened adults, *J. Speech and Hearing Res.*

COMPENSATORY ARTICULATION AND NASAL EMISSION OF AIR IN CLEFT PALATE SPEECH WITH SPECIAL REFERENCE TO THE REINFORCEMENT THEORY

Birgit Hutters 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 primaryly 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. This 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 leads to 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 preference. 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 intelligibili- 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 54 listeners were distributed as follows: 33 mothers, 10 non-educated female cleaners and 11 female school teachers. 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 indivdually presented to each subject and the question was: 'Which of the two pronunciations would you prefer, if you were talking with the speaker that you hear on the tape?'.

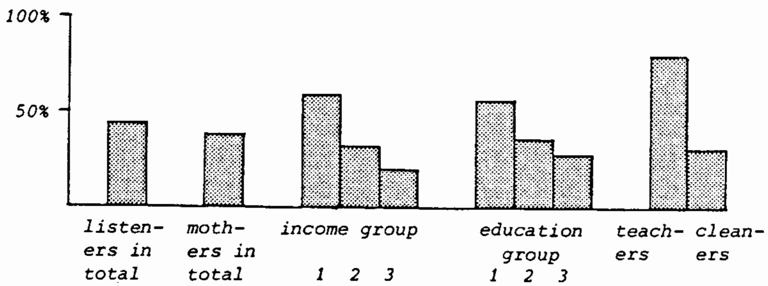
3. RESULTS

In the following, C and E are used for 'compensatory' and 'nasal emission of air', respectively. The results of the C-answers in per cent of the total number of answers 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 between various groups of listeners can be observed. With the group of mothers there are 38% C-answers, but a clear intergroup variation is seen: in the (1) high income group there are 59% C-answers, in the (2) average income group 32%, and in the (3) low 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 rest group (28%), even though only the groups with the highest and the lowest scores are significantly different. 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 score 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



The C-answers in per cent of the total number of answers given by various groups of listeners.

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 material shows only a 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 speech modes. However, after we have listened to the American test tape we think that 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 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 may be 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 we are informed the literature does not report on the relationship between such judgments and the social status and the education of the listeners.

5. REFERENCES

(1) BRADFORD, P.W., CULTON, G.L. (1987), "Parents' perceptual preferences between compensatory articulation and nasal escape of air in children with cleft palate." Cleft Palate Journal 24, 299-303.

(2) BZOCH, K.R. (1979), "Measurement and assessment of categorical aspects of cleft palate speech." In: Bzoch K.R., ed., Communicative disorders related to cleft lip and palate. Boston: Little, Brown.

(3) PAYNTER, E.T, (1987), "Parental and child preference for speech produced by children with velopharyngeal incompetence." Cleft Palate Journal 24, 113-118.

PERCEPTUAL AND ACOUSTIC ANALYSIS OF THE VOICE IN ACUTE LARYNGITIS

Anders Löfqvist and Lucyna Schalén

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 conditions. 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: diplophonia, 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 100-150 ms.

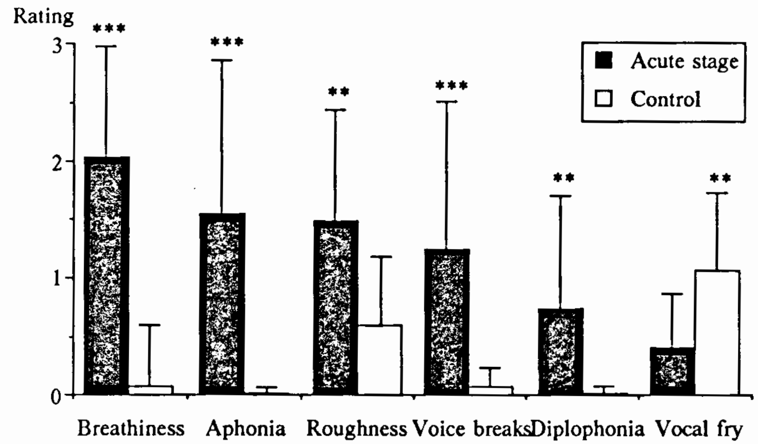


Figure 1. Results of perceptual analysis. *** $p<0.001$; ** $p<0.005$.

For both analyses, a 12-bit A/D conversion was used. The sampling rate was 20 kHz for the long-time spectral analysis, and 10 kHz for the N/S analysis.

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, $t(20) = 4.539$, $p<0.01$. Similar results were found for the relationship between harmonic and non-harmonic 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 parts 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, $t(16) = 2.168$, $p<0.05$. 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 ($r = -0.38$, $p<0.05$), and roughness and N/S ($r = 0.5$, $p<0.01$). The correlations between the

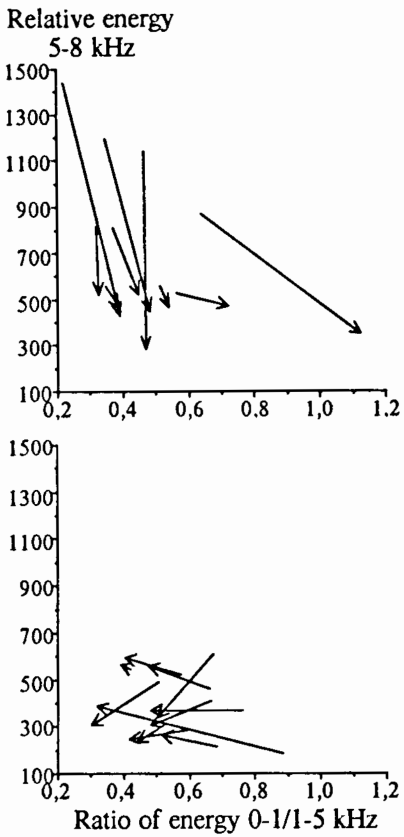


Figure 2. Results of long-time-spectral analysis. The arrows indicate the direction and magnitude of change between the acute and control conditons.

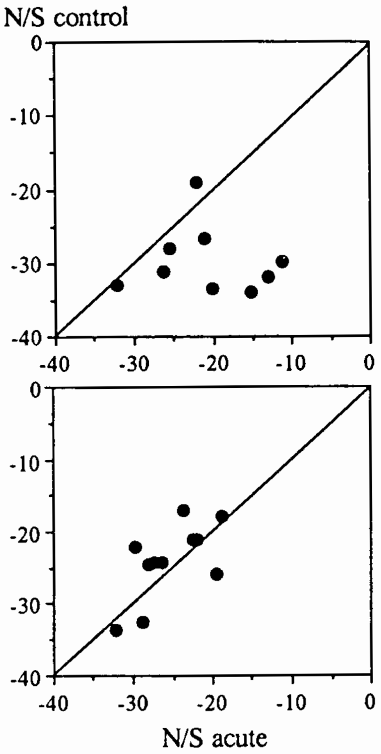


Figure 3. Results of N/S analysis.

ratings of vocal fry and all acoustic measures were negative, although only one was statistically significant.

4. DISCUSSION

The results of the present study indicate that the perceptual ratings of the voices differed between the acute and control conditions. The acoustic analysis did not reveal any overall consistent findings. Rather, two patterns of change were identified.

With the exception of vocal fry, all other perceptual qualities showed a decrease from the acute to the control condition. The changes are most likely related to inflammatory changes in the laryngeal mucosa: edema and a decreased mucosal wave. Vocal fry is presumably a common characteristic of normal voices.

The acoustic analyses suggest two patterns of change between conditions. In one of them, shown in the top panels of Figures 2 and 3, the relative energy between 5-8 kHz decreases as well as the amount of non-harmonic energy. These voices thus contain less noise in the control condition, presumably due to a better glottal closure during phonation. The other group, shown in the bottom panels of Figures 2 and 3, shows a reduction in the measure of spectral tilt.

Signifciant correlations were found between some acoustic and perceptual results. Breathiness was positively correlated with relative energy between 5-8 kHz. This is reasonable, given that this acoustic measure is an indicator of noise. Vocal fry was negatively correlated with relative energy between 5-8 kHz. Again, this is reasonable given that voices characterized by vocal fry can be assumed to contain less noise. Roughness was positively correlated with N/S. Roughness is most likely related to both the amount of noise and to perturbations in time and amplitude. Interestingly, the N/S measure is sensitve to both these aspects of the voice source. That is, a high degree of perturbations will increase the value of N/S.

We should note, furthermore, that the acoustic measures we have applied are related to the frequency domain. The perceptual qualities of voice breaks, aphonia, and diplophonia are most likely related to temporal properties of the voice source. Hence, we should not expect them 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 perceputal 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

[1] ANDERS, L., HOLLIEN, H., HURME, P., SONNINEN, A. & WENDLER, J. (1988), "Perception of hoarseness by several classes of listeners", *Folia Phoniatrica, 40*, 91-100.
[2] GELFER, M. P. (1988), "Perceptual attributes of voice: Development and use of rating scales", *Journal of Voice, 2*, 320-326.
[3] HAMMARBERG, B. & ASKENFELT, A. (1986), "Speech waveform perturbation analysis: A perceptual - acoustical comparison of seven measures" *Journal of Speech and Hearing Research, 29*, 50-64.
[4] HAMMARBERG, B., FRITZELL, B., GAUFFIN, J. & SUNDBERG, J. (1986), "Acoustic and perceptual analysis of vocal dysfunction", *Journal of Phonetics, 14*, 533-547.
[5] HAMMARBERG, B., FRITZELL, B., GAUFFIN, J., SUNDBERG, J. & WEDIN, L. (1980), "Perceptual and acoustic correlates of abnormal voice qualities", *Acta Otolaryngologica, 90*, 441-451.
[6] HAMMARBERG, B., FRITZELL, B. & SCHIRATZKI, H. (1984), "Teflon injection in 16 patients with paralytic dysphonia: Perceptual and acoustic evaluations", *Journal of Speech and Hearing Disorders, 49*, 72-82.
[7] LÖFQVIST, A. & MANDERSSON, B. (1987), "Long-time-average spectrum of speech and voice analysis", *Folia Phoniatrica, 39*, 221-229.
[8] MUTA, H., BAER, T., WAGATSUMA, K., MURAKOA, T. & FUKUDA, H. (1988), "A pitch-synchronous analysis of hoarseness in running speech, *Journal of the Acoustical Society of America, 75*, 224-230.
[9] YANAGIHARA, N. (1967), "Significance of harmonic changes and noise components in hoarseness", *Journal of Speech and Hearing Research, 10*, 531-541.

THE DEVELOPMENT OF ARTICULATORY SKILLS IN CLEFT PALATE BABIES

Kino Jansonius-Schultheiss

Dept. of Phoniatrics, 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 develop–ment (from 0;2-2;0 years of age) was registrated by video recordings of 20 mi-nutes each. Besides that, a larger group of 40 2;0 toddlers (30 CP and 10 normal bornes, 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 pro-cesses than the CP children with an oral plate (13). In comparison with their normal peers, the CP children es–tablished 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, 'under-standibility' 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 the 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. These 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, +/- plosive 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 response; furthermore when the trained transcri–bents 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 (+Cl) and 2 normal born (-Cl) children (Ch).

Ch	1	2	3	4	5
Cl	+	+	+	-	-
la	60	53	56	25	28
ar	12	42	16	19	39
w	20	1	19	44	23
i	8	4	9	12	10

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 (Cl) and 2 normal children (Ch) measured in the period of 0;2 until 2;0 years of age.

Ch	1	2	3	4	5
Cl	+	+	+	-	-
aa	43	50	41	73	68
ap	57	50	59	27	32
ba	44	41	35	58	53
bp	15	14	41	18	20
bv	41	14	41	24	27
w	20	1	19	44	23

In Table 3. all the single articulation movements (tokens), also in babbling, were counted and categorized 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), anterior plosives (Nap) as well as fricatives (Naf), plus the ratio of anterior plosives and fricatives with other articulations (%) in 1 CP (Cl) and 2 normal children (Ch) (from week 10-77).

Ch	3	4	5
Cl	+	-	-
N a	136	591	1054
N at	14	38	40
N ap	2	193	453
N af	3	14	2
% ant. articul.	2	35	43

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 amount of reinforced articulations (Na) and percentages (%): the number of verbally modelled articulation types (Nat), anterior plosives (Nap) as well as frica–tives (Naf), 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).

Mothers of normal born children no.	4	5
N ra	174	142
% ra	29	13
N rat	21	22
% rat	55	55
N ap	71	78
N af	5	2
% ant art.	44	56

Both mothers differend 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 strengthed by the mothers. It looked as if they selected very carefully from all articulatory strings they heard out off 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 languague 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.

Ch	1	2	3	4	5
week	74		80	53	55

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 forst two years of life. The anterior articu– lations 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 effects 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

[1] KOOPMANS-VAN BEINUM, F.J., JANSONIUS-SCHULTHEISS, K. & VAN DER STELT, J.M. (1990), The Influence of an Oral Plate upon the Speech Development and Interaction in the First Years of Life of Cleft Palate Babies, *IFA-report 110*, Institute for Phonetic Sciences, University of Amsterdam, (in Dutch).

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 change around 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 inflected 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 reading 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. Its 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 salience 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 scattergrams, 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 column 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 freqency. More complex descriptions deal with change in the spectrum.

One 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 the channel and its predecessor. This is done for both the basic and the sharpened spectrum, giving two more derived histograms.

A parallel set of histograms is 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 maximum, the value associated with 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 column showed the number of observations at a particular amplitude. In the other, it showed the number of observations which differed from their predecessor by a particular amount (using 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 all inflections, at maxima, and at minima; 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, rise/falls, levels, fall-rises, falls, and compound stretches. The last type contains stretches with more than one inflection. One histogram showed the distribution of these types. A second showed the distribution of stretch durations. A third set out the distribution 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 calculated its mean, variance, and quartile points. Histograms whose x axis was frequency were also described in another way, by summing the values associated with four frequency bands. These were chosen to span the usual range of F0, F1, F2, and frication respectively, using values cited by Baken [1] to set boundaries (which were slightly different for males and females).

3. RESULTS

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 interations 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 1/3 octave in the deafened speakers. The deafened also show too much 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 centre of the spectrum. 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 too sharply 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. The deafened show 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 is higher relative to 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 1: Averages of subjects' median pitch.

	hearing	deafened
females	185.6Hz	199.7Hz
males	119.7Hz	134.8Hz

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 stretches where F0 continues unbroken 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 rather elaborate phrases 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 much less than any other group do, and features which end in a fall much more. It is tempting to link this to the concept of declination as a 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 2: variance of amplitude as a function of hearing and passage.

	read	spontaneous
controls	67	71
deaf	86	79

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 happened.

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 minima. Conversely both variance of amplitude at maxima and variance of amplitude at minima were too great in the deafened. Again, we would expect the opposite if the deafened were simply oscillating between loud and silent.

More detail comes from the properties of the segments 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 behaved rather 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 about the right time.

Combining these observations, only one obvious explanation for the general high variance of amplitude remains. It is that deafened speakers protract relatively extreme events (vowels at one extreme, pauses at the other) for too long. Subjectively this seems true, but it needs direct confirmation.

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 so has 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 correlations 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 promise in the technique of using a battery of statistical descriptors to characterise speech as a distribution of energy, and we are applying it to other domains. 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 drift 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

[1] BAKEN, R. (1987), *"Clinical measurement of speech and voice"*, London: Taylor & Francis.

[2] GOEHL, H. & KAUFMAN, D. (1984), "Do the effects of adventitious deafness include disordered speech?", *Journal of Speech and Hearing Disorders*, *49*, 58-64.

MEAN-TERM PERTURBATIONS OF THE PSEUDO-PERIOD OF THE GLOTTAL WAVEFORM

R. De Guchteneere and J. Schoentgen*

Institute of Phonetics, Free University of Brussels
Av. F.D. Roosevelt, 50, B-1050 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 analysis interval of, typically, fifty periods. Frequently, the differences between individual periods and a local 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 volitional 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 consist of supposing that any systematic relationship between neighbouring periods can be effectively removed by smoothing and that, moreover, what is so removed is not worth 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 term fluctuations appear to be superimposed on stochastic mean-term perturbations. These do 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, 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], [i], and [u]), 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 cms from the lips. The laryngograph (or EGG for electroglottograph) 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 redigitized 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.

The algorithm that we 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 centred on the main events was oversampled eight times and low-pass filtered; the period markers were then detected with improved accuracy, leading to a theoretical resolution in time of 6.25 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 on most of 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 average, 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 a great 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 five male and Figure 1b the time series of five female speakers. The first sequence in figure 1b presents a case of a negative correlation between neighbouring periods; all the other sequences present positive correlations.

The mechanisms underlying the production of jitter are not yet fully understood. Neurological and cardiac mechanisms, which have been shown to contribute to jitter [1], [5], would lead us to expect perturbations of the fundamental period straddling several cycles. Indeed, in an enumeration of possible candidate 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.

	Speech signal						EGG signal					
	Pearson			Rank			Pearson			Rank		
	+	0	-	+	0	-	+	0	-	+	0	-
[a]												
Healthy sp. (13)	10	1	2	12	0	1	10	1	2	11	1	1
Dysphonic sp. (12)	9	2	1	11	0	1	9	0	3	11	1	0
TOTAL (25)	**19**	**3**	**3**	**23**	**0**	**2**	**19**	**1**	**5**	**22**	**2**	**1**
[i]												
Healthy sp. (13)	5	3	5	13	0	0	5	3	5	13	0	0
Dysphonic sp. (12)	9	1	2	11	1	0	10	1	1	12	0	0
TOTAL (25)	**14**	**4**	**7**	**24**	**1**	**0**	**15**	**4**	**6**	**25**	**0**	**0**
[u]												
Healthy sp. (13)	11	2	0	13	0	0	11	2	0	13	0	0
Dysphonic sp. (12)	8	1	3	10	1	1	10	0	2	12	0	0
TOTAL (25)	**19**	**3**	**3**	**23**	**1**	**1**	**21**	**2**	**2**	**25**	**0**	**0**

4. REFERENCES

[1] Baer, T. (1980) : Vocal jitter : A neuromuscular explanation. Transcripts of the Eighth Symposium of the Care of the Professional Voice, Voice Foundation, New-York, pp 19-22.

[2] Hess, W. and Indefrey, H. (1987) : Accurate time-domain pitch determination of speech signal by means of a laryngograph, Speech Comm., 6, pp 55-68.

[3] Kasuya, K., Kobayashi, Y., Kobayashi, T., Ebihara, S. (1983) : Characteristics of pitch period and amplitude perturbations in pathologic voice, IEEE-ASSP Conference, Boston, pp 1372-1375.

[4] Lieberman, Ph. (1963) : some acoustic measures of the fundamental periodicity of normal and pathologic larynges, J. Acoust. Soc. Am., 35, 3, pp 344-353.

[5] Orlikoff, R.F. and Baken, R.J. (1989) : Fundamental frequency modulation by the heartbeat: preliminary results and possible mechanisms, J. Acoust. Soc. Am., 85, pp 888-893.

[6] Pinto, N.B. and Titze I.R. (1990) : Unification of perturbation measures in speech signals, J. Acoust. Soc. Am., 87, 3 , pp 1278-1289.

[7] Schoentgen, J. and De Guchteneere R. (1990) : An algorithm for the measurement of jitter, Proceedings of the ESCA Workshop on Speaker Characterization, Edinburgh, pp 175-180.

[8] Titze, I.R., Horii, Y., Scherer, R.C. (1987) : Some technical considerations in voice perturbation measurements, J. Speech Hear. Res., 30, pp 252-260.

Fig. 1a

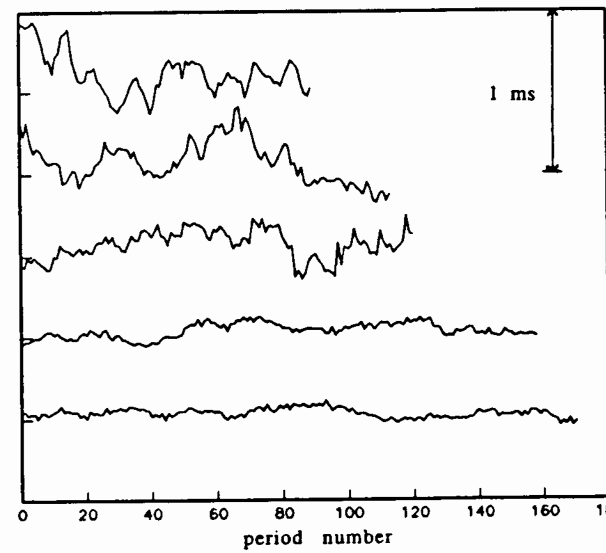


Fig. 1b

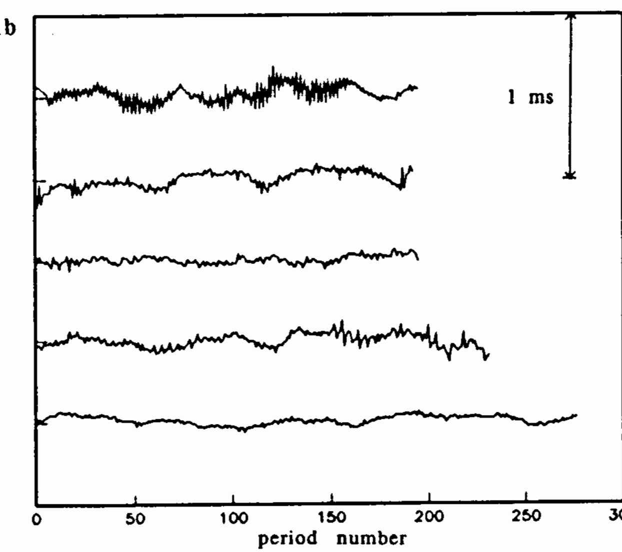


Figure 1

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 neighbouring 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 performance 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 latters' 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 [2]. 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 perfom 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 subtests 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.

500 normal hearing children (ages between 3 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.

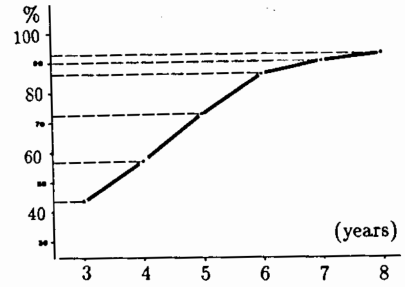


Figure 1.
Performance of children

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 reexamined 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 Assesment 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

6 from Class B) who had both problems: metatheses and disturbed directions. 4 children could not correctly repeat rhythmic sentences. 5 boys and 3 girls of the total 37 had articulation problems (generally mispronunciation of sibilants). The majority of children were right-handers: 21 of the two classes, while 8 (5 from Class A and 3 from Class B) were left-handers and another 8 children had no dominant hand (6 of them used their right hands for drawing and eating).

Table 1

Results of speech perception/ comprehension examinations

GMP-subtests (examinations)	Children's perform. Class A 1st	2nd
lip-reading	40%	50%
word-completion	3.8	4.5
visual memory	5.6	5.6
verbal memory	4.7	4.7
nonsense words	84.1%	95%
speeded-up sent.s	71.2%	90%
noisy sent.s	88.2%	100%
noisy words	88.8%	100%
filtered sent.s	100%	100%
natural sent.s	100%	100%
text-compr.	60%	80%
Average	79%	89.3%

GMP-subtests (examinations)	Children's perform. Class B 1st	2nd
lip-reading	28%	30%
word-completion	3.6	4.0
visual memory	5.6	5.6
verbal memory	4.5	4.5
nonsense words	86%	90%
speeded-up sent.s	65.3%	70%
noisy sent.s	86.5%	90%
noisy words	83.4%	90%
filtered sent.s	100%	100%
natural sent.s	100%	100%
text-compr.	53.5%	70%
Average	75.3%	80%

The children's data show various co-occurrences of problems as shown by the GMP-subtests, such as a mixed-hander pronouncing metatheses, having problems in identifying the speeded-up sentences, or a riht-hander 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. On the basis 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.

For the sake of the experiment, the children's GMP results were disclosed only to one of the two teachers, the one who taught in Class A. Moreover, some corrective excercises were proposed to this teacher to be used in the classroom in order to: (i) stabilize the children's directions and hand dominance (where this was necessary), (ii) improve their speech perception performance and general language skill, and (iii) extend their own vocabulary. The results of the re-examination 4 months later confirmed the usefulness of these corrective excercises in teaching reading. The children's performance in a Reading Assessment Test at the end of the 4-month period supported our hypothesis referred to above. This test contains 6 subtests: a letter identification task, word reading controlled by pictures, words containing a missing

letter, isolated sentence understanding controlled by a drawing task, reading text comprehension controlled by questions for words and sentences. The maximum score was: 100 points. Table 2 shows the data of the Reading Assesment Test.

Table 2

Interrelation of the children's GMP results and their reading performance

Classes GMP (1st/2nd)	Average performance in reading test	understanding of reading
A 79/89.3%	97.41 points	93.5 points
B 75.3/80%	87.5 points	79.2 points

The children's performance with the GMP test-package shows significant difference 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 level 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

Predictions and supporting data on reading ability

Predicted readers	Average GMP results (%)	Perform. in RAT (points)
'good'	88.6	95-100
'poor'	65.1	90-96*
'poor'	66.3	65-85**

* (after corrective course)
** (without corrective course)

Table 4

Distribution of children according to their results in reading test

Points	Distribution of children according to RAT results (%) Class A	Class B
100	53.1	35
95-99	29.5	10
90-94	17.4	20
85-89	–	5
80-84	–	10
75-79	–	5
70-74	–	10
65-69	–	5

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 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 a preschool age. This offers a good prognosis for successful reading acquisition.

6. REFERENCES

[1] GÓSY, M. (1989), *"Beszédészlelés/ Speech perception"*. Budapest: MTA.
[2] LIEBERMAN, Ph. et al. (1985), "Phonetic speech perception deficits in dyslexia", *JSHR 28*, 480-486.
[3] SHANKWEILER, D., CRAIN, S. (1986), "Language mechanisms and reading disorder: A modular approach", *Haskins Laboratories, SR*, 173-197.

ANALYSIS OF GLOTTAL WAVEFORM IN DIFFERENT PHONATION TYPES USING THE NEW IAIF-METHOD

P. Alku[1], E. Vilkman[2], U.K. Laine[1]

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 fourty years in order to estimate the source of voiced speech, the glottal pulseform. One of the most popular techniques that is applied 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 AIF-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

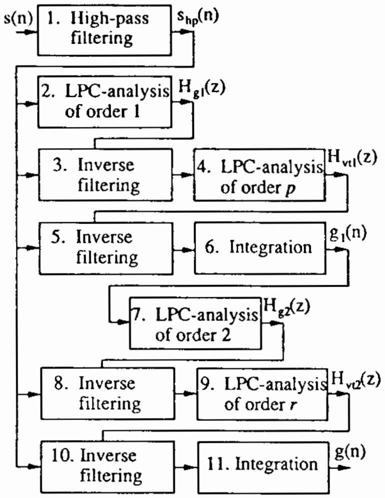
The IAIF-method is based on a speech production model that consists of three separated processes: the glottal excitation, the vocal tract and the lip radiation effect. The model is assumed to be linear and the interaction between the three parts is considered to be negligible. The vocal tract is modeled 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. 1:
In order to remove undesirable fluctuations of the output of the integrator signal s(n) has to be high-pass filtered. The high-pass filter is a linear phase FIR having 511 coefficients and a cut-off frequency of 20 Hz.

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 $H_{g1}(z)$.
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 estimated is $g_1(n)$.
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 $H_{vt2}(z)$.
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.



Transfer functions of the filters are:

$H_{g1}(z) = 1 + az^{-1}$ $\quad$ $H_{vt1}(z) = 1 + \sum_{k=1}^{p} a(k)z^{-k}$

$H_{g2}(z) = 1 + bz^{-1} + cz^{-2}$ $\quad$ $H_{vt2}(z) = 1 + \sum_{k=1}^{r} b(k)z^{-k}$

Fig. 1. Block diagram of the IAIF-method

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.

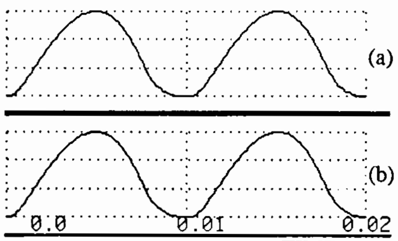


Fig. 2. Analysis of a synthetic male vowel of breathy phonation
(a): Original synthetic glottal source
(b): Glottal wave estimate given by the IAIF-method (p=8)

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 of Fig. 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.

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 (Brüel&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 downsampled 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.

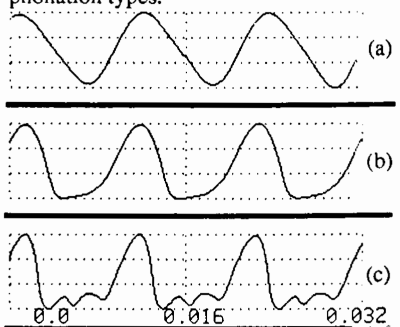


Fig. 3. Glottal wave estimates given by the IAIF-method (natural male voice, p=12)
(a): Breathy phonation
(b): Normal phonation
(c): Pressed phonation

The analysis of female voice yielded results that were, quite surprisingly, free from formant ripple for all the three phonation types. The waveform of breathy phonation was of a very smooth shape. No clear closed phase could be distinguished. The time instant of the maximum glottal opening occurred approximately in the middle of the glottal cycle. In the case of normal phonation the time instant of the maximum opening was moved to the point that corresponds to 70 % of the length of the pitch period. The waveform of pressed phonation was the only one with a clear closed phase.

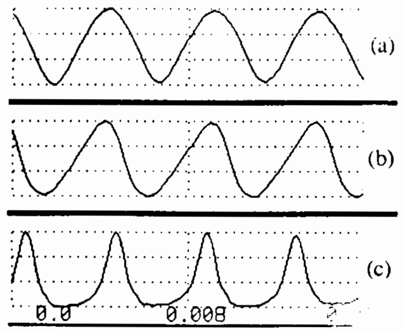


Fig. 4. Glottal wave estimates given by the IAIF-method (natural female voice, p=12)
(a): Breathy phonation
(b): Normal phonation
(c): Pressed phonation

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 well 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 connected speech. Also the real-time implementation of the algorithm using the TMS320C30-signal processor is under development.

References:

[1] Alku, P., Vilkman, E., Laine, U.K. (1990), "A comparison of EGG and a new inverse filtering method in phonation change from breathy to normal," Proc. Int. Conf. on Spoken Language Processing, pp. 197-200.

[2] Ananthapadmanabha, T.V. (1984), "Acoustic analysis of voice source dynamics," STL-QPSR 2-3, pp.1-24.

[3] Gauffin, J., Sundberg, J. (1989) "Spectral correlates of glottal voice source waveform characteristics," J. Speech and Hearing Research, Vol. 32, pp. 556-565.

[4] Gold, B., Rabiner, L.R. (1968), "Analysis of digital and analog formant synthesizers," IEEE Trans. Audio and Electroacoustics, Vol. 16, pp. 81-94.

[5] Wong, D.Y., Markel, J.D., Gray, A.H., Jr. (1979), "Least squares glottal inverse filtering from acoustic speech waveforms," IEEE Trans. Acoust., Speech, Signal Processing, Vol. 27, pp. 350-355.

A VOICE CONVERSION METHOD AND ITS APPLICATION TO PATHOLOGICAL VOICES

H. Kuwabara[1] and T. Takagi[2]

1 The Nishi-Tokyo University, Yamanashi, Japan
2 NHK 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 trajectory has the largest 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 trajectory patterns.

Based on the experimental evidence mentioned above, an experiment has been performed to change and improve the quality of natural speech making use of the analysis-synthesis system.

Formant trajectories are extracted first from voiced portions by LPC method and the dynamics of these trajectories are altered depending on the formant pattern itself. The method for altering the formant pattern is the same as that we have proposed earlier for the normalization of vowels in connected speech [3]. This method is applied to the formant and pitch trajectories extracted from natural speech, and the quality-controlled speech sounds are synthesized using the analysis-synthesis system to present to listeners for perceptual judgment.

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 signals. Formant frequencies and their bandwidths were estimated by solving a polynomial equation. A modification of the spectral envelope is equivalent to a manipulation of the coefficients that would result in a frequency response of the filter equal to the modified envelope.

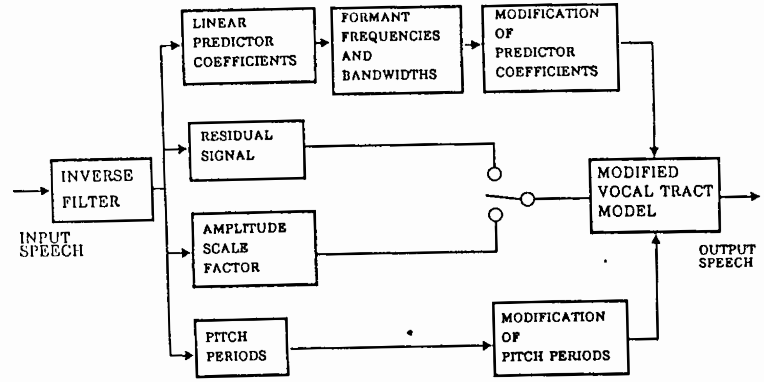


Fig.1 Block diagram of analysis-synthesis system for voice conversion

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 them was conducted in such a way that the preceding and succeeding acoustic features contributed to the present value with the same weight if the time differences from the present were equal, and that the amount of contribution was proportional to the difference from the present acoustic feature [3]. Suppose $x(t)$ be the time-varying pattern of a formant frequency, the new value $y(t)$ is defined as the sum of the original value $x(t)$ and the additional term of contribution by contextual information. The contribution is assumed to be a weighted sum of differences between values at the present time t and at different time $t\pm\tau$

Thus, $y(t)$ is given by

$$y(t)=x(t)+\int_{-T}^{T}w(\tau)(x(t)-x(t+\tau))d\tau \quad (1)$$

where $w(\tau)$ is the weighting function which is given as

$$w(\tau)=\alpha\cdot\exp(-\tau^2/2\sigma^2). \quad (2)$$

In this study, T=150ms and σ=52 ms were experimentally decided. Given $\alpha>0$, the dynamics of the original formant trajectory is emphasized, while for $\alpha<0$, it becomes deemphasized.

Equation (1) is applied to each of the three formant trajectories without vowel/consonant (except for voiceless consonant) distinction. The time interval in equation (1) during which the weighted sum is calculated is 300ms, a 150ms forward and backward each. This is the result for α = 7.3 which, in our previous study, represents a proper value for the purpose of normalizing coarticu-

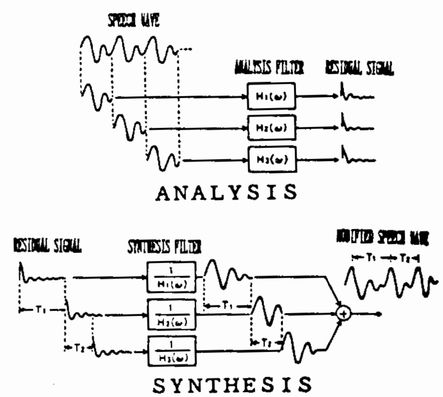


Fig.2 Schematic illustration for changing pitch frequency

lation effects of vowels in continuous speech. It is noticed from the figure that the new formant trajectories are emphasized their up-and-down dynamic movement as compared 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 are eliminated and to lower the frequency, zero signals are added to the last part of the residue.

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 a voice uttered by 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 at first detected and the spectral envelopes are extracted by LPC analysis. Next, 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 devocalized 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

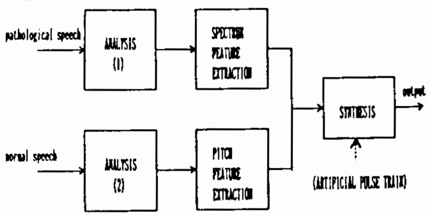


Fig.3 Block diagram for the enhancement of pathological speech

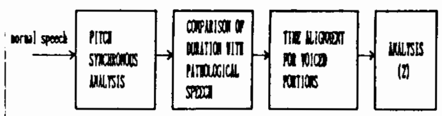


Fig.4 Block diagram of analysis and time alignment for normal speech

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 waveforms for the pathological speech ,synthesized speech 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 is has been found to be far better than the original pathological speech, though it is not as good as the normal speech sound.

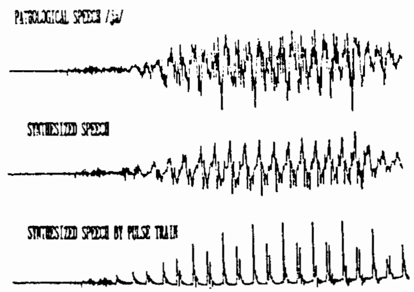


Fig.5 An example of speech waveforms of the pathological and synthesized speech

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 closely relates 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 also 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

[1] H. Kuwabara, "A pitch synchronous analysis / synthesis system to independently modify formant frequencies and bandwidth for voiced speech,"SPEECH COMMUNICATION, Vol. 3 (1984) pp.211-220
[2] H. Kuwabara, K. Ohgushi, "Acoustic characteristics of professional male announcers' speech sounds," ACUSTICA, Vol.55 (1984) PP.233-240
[3] H. Kuwabara, "An approach to normalization of coarticulation effects for vowels in connected speech," J. Acoust. Soc. Amer., Vol. 77 (1985) pp.686-694
[4] H. Kasuya, "An algorithm to choose formant frequencies obtained by linear prediction analysis method," Trans. IECE Japan, Vol. J66-A (1983) pp.1144-1145

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/'s is a low third formant (F3) which is close in frequency to the second formant (F2). F3 for /r/ is usually around 2000 Hz or below, whereas for other segments, F3 is usually above 2400 Hz. There is substantial downward movement in F3 from a canonically articulated /ɑ/ to an /r/ in words like "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 anticipation of /r/.

2. CORPUS

To conduct this acoustic study, the words "car," "cart," "carve," "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" (/Nədɑv/) 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 /ar/ 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 always show durational differences.) As expected, average /ar/ 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

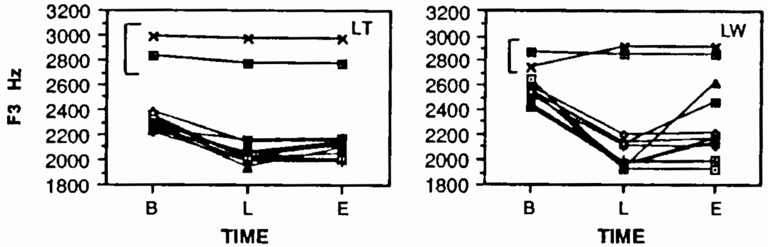


Figure 2. A comparison of F3 trajectories occurring for subjects LT and LW

F3 trajectory for MR always moves downward toward the end of the sonorant region (L-like) when the final consonant is labial, and always moves upwards towards the end of the sonorant region (J-like or U-like) when the final consonant is alveolar. On the other hand, LW shows a rate effect. All of the F3 trajectories show downward and upward movement when speaking slow and only downward movement when speaking fast.

As noted above, all of these patterns are consistent with a theory that the /r/ has a stable trajectory, but variable timing. The implication is that a U-shaped trajectory is always present but not visible. To support the theory, we compare in Figure 3 the /t/ bursts in the slow and fast pronunciations of "cart" by subject LW. The major spectral prominence of the /t/ burst in the fast pronunciation is around 1500 Hz lower than it is in the slow pronunciation. This substantial spectral difference suggests that the /r/ in the fast pronunciation of "cart" is still being articulated during the following /t/.

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 (Fn) during the neutral case,the /ɑ/ 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 Fn. 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 Fn at the beginningof the sonorant region, the normalized difference

Table 1. Shapes of F3 trajectories across all speakers as a function of rate and context. The words are specified by the final consonant.

Shape of F3 Trajectories

	Subjects	L-like	Flat	J-like	U-like
Slow Repetitions	AF	P			V,T,D
	LW				all
	MR	V,P			T,D
	MH	all			
	LT		all		
	JR	all			
Fast Repetitions	AF	P			V,T,D
	LW	all			
	MR	V,P		T	D
	MH	all			
	LT		all		
	JR	all			

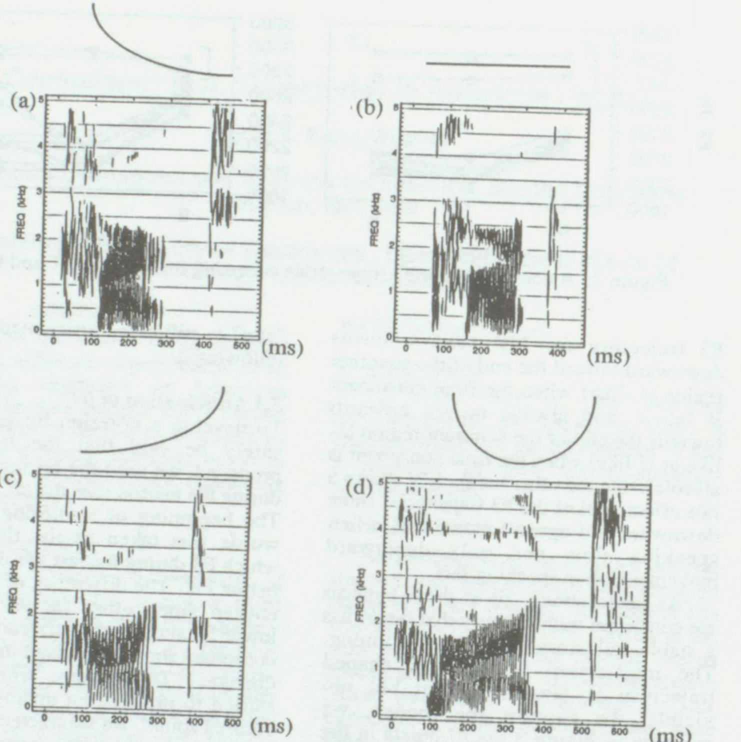


Figure 1. A comparison of F3 trajectories during different pronunciations of "cart."

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 /ɑ/ in the fast and slow pronunciations of "Nadav," 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

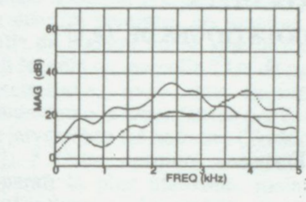


Figure 3. A comparison of the /t/ burtsts in the slow (dotted) and fast (solid) pronunciations of "cart" by subject LW.

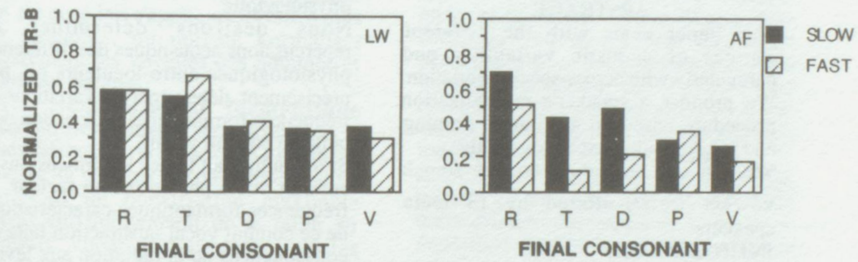


Figure 4. A plot of the normalized difference TR-B for subjects LW and AF

will be 0. If F3 falls 500 Hz below Fn at the end of the sonorant region, the normalized difference will be 1. Values in between indicate where in the sonorant region speakers started to produce the /r/.

A bar graph of this normalized difference is shown in Figure 4 for subjects LW (left) and AF (right). Recall that LW shows two distinct patterns of F3 trajectories when speaking slow and fast. The near equivalence of the column heights for subject LW means that there is little difference in when F3 begins to lower, relative to the beginning of the vowel. Thus, one explanation for the difference in the F3 trajectories is that the articulation of /r/ requires a minimum time of execution so that the upward movement in F3 away from the /r/ gesture is not seen in the case of the fast pronunciations because the sonorant region is too short. In fast speech, then, LW starts to produce the final consonant before finishing the production of /r/, as was seen in Figure 3. Interestingly, data from subject AF, whose F3 lowers at a more variable point in the word, shows that some speakers generally start to produce the /r/ much earlier when 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 speech signal. The sonorant regions tend to change in duration due to factors such as speaking rate and the voicing of any final consonant. Furthermore, what portion of the F3 trajectory occurs during the sonorant region depends on how early speakers begin to produce the /r/. This study has shown that different people have different timing for an /r/ and that these tendencies can vary depending on speaking rate, and the voicing and place of articulation of a following consonant.

5. ACKNOWLEDGEMENTS

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

[1] Espy-Wilson, C. (1987), "An acoustic-phonetic approach to speech recognition: application to the semivowels," RLE Tech. Rep. 531, MIT.

LA VARIABILITE INTER-LOCUTEUR, ETUDE SUR LES REALISATIONS ACOUSTIQUES DE / e, ɛ/

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,ɛ/) uttered by 13 male speakers.

INTRODUCTION

Nous présentons ici une étude des variations qui agissent simultanèment 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 articulatoire. 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 l'analyse de deux voyelles antérieures du français, /e/ et / ɛ /, pour deux principales raisons: d'une part / ɛ/ constitue une bonne voyelle d'apprentissage et d'autre part il nous semble particulièrement intéressant d'étudier deux voyelles dont l'opposition est neutralisable en français.

1. VARIATION INTER-LOCUTEUR

Nous tenterons d'évaluer les effets de la variation inter-locuteur d'origine physiologique et nous évoquerons les conséquences de la variation inter-locuteur d'origine articulatoire.

1.1 Tentative d'évaluation et de normalisation de la variation d'origine physiologique

Nous désirons déterminer les répercussions acoustiques des différences physiologiques entre locuteurs ou, plus précisèment, 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 un 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:
Ff/Fh 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 (Fant [2]) et l'ouverture relative au point de constriction maximale (Traünmuller [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 mue, causée par la baisse du larynx chez les hommes (Traünmuller [3]) ? Cette dernière solution nous apparaît la plus plausible, mais nous analyserons néanmoins les deux éventualités. Remarquons auparavant qu'au-delà de cette interrogation d'ordre physiologique, c'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 séparèment les voyelles 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 linéairement en fonction de cette taille et que les différences physiologiques constituent l'unique source de variation fréquentielle. Cette variation peut être simplement évaluée par la position relative des fréquences formantiques d'un locuteur par rapport aux fréquences 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 pourrait donc 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 et dont les conséquence 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 perpendiculaire d'une image acoustique d'une voyelle V sur la droite "physiologique" de V et à se reporter au critère de la position relative, en le modifiant légèrement puisque deux formants au moins sont désormais nécessaires pour notre évaluation. Cette méthode, proposée par F 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.
A partir des fréquences formantiques d'une voyelle V prononcée par un locuteur l ainsi que des fréquences formantiques moyennes des hommes et des femmes pour cette même voyelle V, nous déterminons un paramètre de normalisation, appelons-le "dist", qui tient compte implicitement de la taille du conduit vocal de l. Détaillons la procédure:

-nous enregistrons trois répétitions de / ɛ / en contexte labial, émises par un locuteur l. En théorie, une seule répétition de /ɛ/ suffit mais trois répétitions au moins sont nécessaires afin de minimiser les erreurs dues à une prononciation déviante. /ɛ/ nous semble être une bonne voyelle d'apprentissage (Bonneau [4])

- nous projetons l'image acoustique de /ɛ/ -la moyenne des trois répétitions- sur la droite qui relie les moyennes masculines (Fhɛ) et féminines (Ffɛ) de /ɛ /.
-nous calculons le paramètre de normalisation physiologique "dist" qui indique la position relative de notre projection par rapport à Ffɛ et Fhɛ.

-toutes les voyelles de /ɛ/ émises par l sont normalisées par l'application de la

formule:
soit Flij représentant la fréquence du ième formant de la voyelle j prononcée par le locuteur l et F'lij sa normalisation,
F'lij=Flij - dist * FhijFfij.
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 plus 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 un 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 à nouveau 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

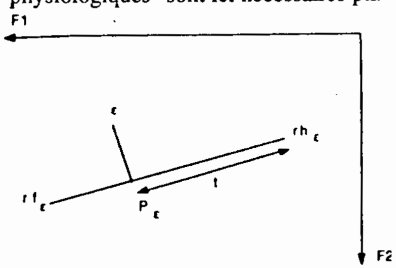


Figure 1
Détermination du paramètre de normalisation t,

rhε moyenne masculine de ε

rfε moyenne feminine de ε

ε: représentation acoustique de cette voyelle pour le locuteur sur lequel on effectue la normalisation, Pε 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 réevaluer 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 articulatoire 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 /u/, qui peut remettre en cause la validité des droites "physiologiques",
- les variations fréquentielles entre locuteurs d'origine articulatoire, si elles ne sont pas négligeables 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 normalisons-nous réellement, la variation d'origine physiologique, articulatoire ou une partie des deux?

2. METHODOLOGIE

Le corpus d'évaluation est constitué de deux phrases qui comportent les réalisations des deux voyelles / ε, e / ,une voyelle par phrase, en syllabe accentuée et dans des contextes consonantiques symétriques : labial, dental, palatal et uvulaire. Voici ces phrases, nous avons séparé les voyelles étudiées par un espace.

"A Papeete, son fr è re a mangé cette f è ve faite en li è ge"
"Vous ser e z raccord é si vous pouv e z pay e r chaque mois"

/ ε / et / 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 / ε / et syllabe ouverte suivie d'une frontière morphologique pour / e / . Treize locuteurs masculins ont enregistré le corpus, que nous espèrons compléter par les données d'autres locuteurs et surtout d'autres locutrices pour la présentation de ce papier 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, ε / est respectée puisque les réalisations acoustiques de ces deux voyelles sont bien distinctes dans un contexte donné. D' autre part l'effet de chaque contexte consonantique est parfaitement prévisible quel que soit 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 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%) ,70% (76%) pour F1, F2, F3 de /ε/,
– 5%,51%,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 séparèment, à une exception près: le premier formant de / ε / en contexte palatal, 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 / ε / 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 resultats obtenus par voyelle et par formant semblent satisfaisants si on considère que seules des voyelles émises par des locuteurs masculins ont pour l'instant été normalisées. L' emploi de barks à la place des Hertz améliore légèrement les résultats obtenus pour F1 en rééquilibrant pour chaque formant les écarts fréquentiels entre les moyennes ou références acoustiques.
L'articulation spécifique à chaque locuteur varie avec le contexte et s'exprime le long de nos droites "physiologiques" ce qui pertube les résultats obtenus pour F2 / ε /. Il est difficile de commenter les résultats obtenus pour F1 /e/, la variation inter-locuteur étant relativement faible avant la normalisation.

4. REFERENCES

[1] PETERSON, G.E. and BARNEY, H.L (1952), "Control Methods Used in a Study of the Vowels",JASA, vol 24, pp175-184.
[2] FANT, G. (1973), "A note on vocal tract size factors and nonuniform F-Pattern scalings", Speech Sounds and Features, MIT Cambridge, USA, pp 84-93.
[3] TRAUNMÜLLER, H. (1984), "Articulatory and perceptual factors controlling the age- and sex-conditioned variability in formant frequencies of vowels", Speech Communication, vol 3, numéro 1, pp 49-62.
[4] BONNEAU, A., FOHR, D., LONCHAMP, F. (1989), "Normalisation formantique des locuteurs féminins et masculins à l'aide de connaissances et d'un apprentissage minimal" Actes du séminaire sur la variabilité du locuteur, Luminy, juin 1989.

SOME SARA VOWEL INVENTORIES AND VOWEL SYSTEM PREDICTIONS

D. I. Djarangar

Institut de la Communication Parlée
URA CNRS n°368 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 [ɛ] which, according to Crothers [1], should appear "earlier" than phonemic [ɨ] and [ə]. Still, in Sara languages, [ɛ] 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 Moïssala; 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 [...tVt...] 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: [tí:] *it swelled*, [té:] *he deceived him*, [tɛ́:] *he married her many times*, [tá:] *he took it many times*, [tɔ́:] *he tied up*, [tó:] *he blew on it*, [tú:] *he swallowed* , [tɨ́] *to, into*, [tə́:] *he deceived*, [tǿ:] *he blew on him*, [tœ́:] *he tied him*. But in underlying phonological representations, only [i, e, a, ɔ, o, u] are phonemic. Actually, [ə] and [ɨ] are, respectively, allophones of [e] and [i]. Furthermore [ɛ:, œ:, ø:] are respectively phonotactic fusions of [a+e], [ɔ+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 Adaptative 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, ɔ, 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 [ə]. In Kaba (fig. 4), the central high vowels [ɨ] and [ə] 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 [ø] and [ə] in Bediondo Bedjond, which are two very close allophones of different phonological vowels (cf. [8] for a study of this special case, [ø] like [œ], being clearly *rounded*).

3. VOWEL INVENTORIES AND THE LACK OF PHONEMIC [ɛ]

According to [1], languages with five or more vowels have phonemic [ɛ]. Most of the time, the five vowels are [i, e, a, ɔ, u]. The prediction of [ɛ] 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, ɔ, o, u, ɨ]. According to the descriptions of [8] and [10], [ɨ] is an allophonic variant of [i]. It is interesting to notice that while Sar attests [ɨ], it does not have [ɛ], although, from a typological point of view, one could expect an [ɛ] first. The acoustic vocalic triangle clearly shows here the gap left by this absence of [ɛ] (fig. 1).

Mbay has 8 vowels: 6 peripheral [i, e, a, ɔ, o, u] and 2 interior [ɨ, ə] [9]. Mbay does not attest [ɛ]. The gap left by its absence appears also clearly in the acoustic space (fig. 2).

Nar also has 8 vowels. It attests [ɛ] but only at a *phonetic* level: [ɛ] is a result of [e] in a preconsonantal position (fig. 3).

In Bedjond dialects [6], [ɛ:] (always realized long) is a combination of a final [...a#] with the pronoun [-e] "his" (e.g. [tà:] *mouth* + [é] *his* → [tɛ̌:] *his mouth*).

Only Kaba seems to have an [ɛ] 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.

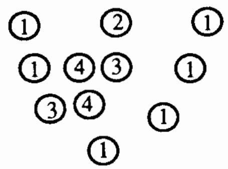
4. THE ORDER OF APPEARANCE OF INTERIOR VOWELS IN SARA LANGUAGES

According to [1], six-vowel systems have [i, e, ɛ, ɔ, o, u] or [i, ɛ, a, ɔ, o, ɨ]. The second system is said to be the most common. Generally, when languages have 9 vowels, there are 7 peripheral vowels [i, e, ɛ, a, u, o, ɔ]. The other 2 vowels are the front rounded [y, ø], the back unrounded [ɯ, ɤ], or central [ɨ, ə].

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 [ɛ], it seems easier to develop an [ə] after an [ɨ] (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 [ø, œ].
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 ①,



the tendency is first to develop allophones at the high central position, ② (Sar); before filling the remaining cardinal (Nar and Beda Bedjond) or central (Mbay) positions, ③ (both, for Kaba); then, to exploit the front rounded positions, ④ (Bediondo Bedjond).

5. SARA LANGUAGES AND VOWEL SYSTEM PREDICTIONS

Again according to [1], "Languages with six or more vowels have ɔ and also either ɨ or e , generally the former" and "Languages with seven or more vowels have e, o or ɨ, ə ".
Sara languages have 6 *phonemic* vowels. In accordance with Crothers' prediction, they have [ɔ] and [e], but they do not have phonemic [ɨ] and [ə]. But at a *phonetic* level, all of them have [ɨ], and 4 of them also have [ə] (including [e] = [ə] of Beda 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 [ɛ].

ACKNOWLEDGEMENTS

Many thanks to C. Abry for his fruitful suggestions and T. Sawallis for improving our English.

REFERENCES

[1] CROTHERS, J. (1978) "Typology and universals of vowel systems", *Universals of Human Language*, J.H. GREENBERG ed., Stanford: Stanford University Press, vol. 2: 95-125.
[3] CAPRILE, J.P. (1977) "Quelques problèmes de phonologie en mbay de Moïssala", *Etudes phonologiques tchadiennes*, J.P. CAPRILE éd., Paris: SELAF, 22-35.
[6] DJARANGAR, D. (1989) *Description phonologique et grammaticale du bédjonde*, Thèse de doctorat, Grenoble 3.
[8] _______ (1991) "Etude acoustique du système vocalique du sar", *Proceedings of the 4th Nilo-Saharan Linguistics Colloquium*, L. BENDER ed., Hamburg: Helmut Buske Verlag, 295-314.
[9] _______ (1991) "Analyse acoustique et interprétation des voyelles centrales du mbay", *Linguistique Africaine*, 6, Paris: GERLA
[5] GREENBERG, J.H. (1963) *Languages of Africa*, The Hague: Mouton & Co.
[2] LILJENCRANTS, J. & LINDBLOM, B. (1972) "Numerical simulation of vowel quality systems: the role of perceptual contrast", *Language* 48: 839-862.
[3] LINDBLOM, B. & ENGSTRAND, O. (1989), "In what sense is speech quantal ?", *Journal of Phonetics* 17: 107-121.
[4] MADDIESON, I. (1986) *Patterns of sounds*, Cambridge: Cambridge University Press.
[10] PALAYER, P. (1990) *La langue sar*, Thèse de doctorat, Université de Tours.

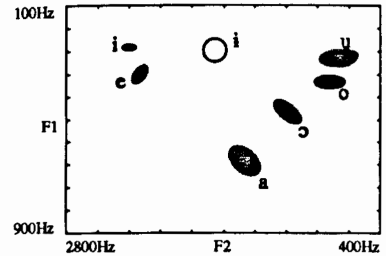


Figure 1: Sar

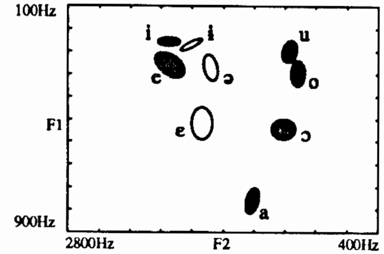


Figure 4: Kaba

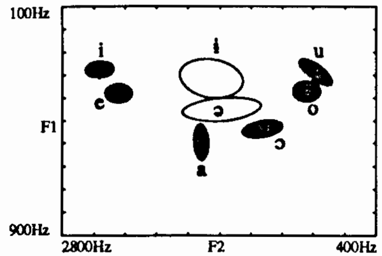


Figure 2: Mbay

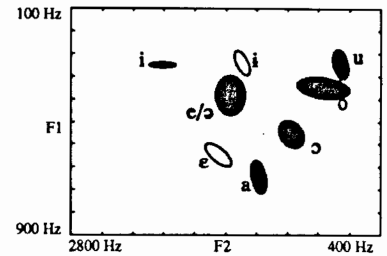


Figure 5: Beda Bedjond

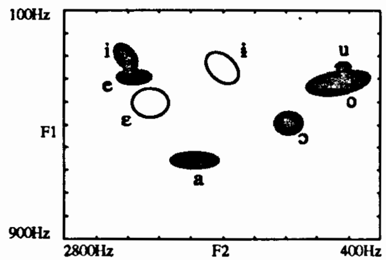


Figure 3: Nar

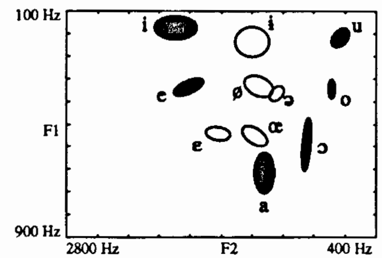


Figure 6: Bediondo Bedjond

Formant mean values

		i	e	ɛ	a	ɔ	o	u	ɨ	ə	ø	œ
Sar	F1	252	346		651	483	370	290	261			
	F2	2331	2250		1463	1120	798	723	1681			
Mbay	F1	328	411		580	535	401	336	359	467		
	F2	2520	2375		1746	1271	931	863	1664	1591		
Nar	F1	273	336	431	631	496	365	303	310			
	F2	2342	2286	2158	1821	1104	733	695	1621			
Kaba	F1	235	314	519	796	540	340	269	243	328		
	F2	2031	2010	1776	1393	1150	1036	1088	1848	1699		
Beda Bedjond	F1	305	401	631	700	558	386	305	305			
	F2	2091	1565	1648	1340	1086	853	721	1480			
Bediondo Bedjond	F1	168	375	546	680	581	386	201	214	400	375	550
	F2	1968	1874	1641	1281	950	765	683	1376	1198	1305	1346

LOCUS-NUCLEUS RELATION AND VOT IN SPONTANEOUS AND ELICITED SPEECH

Diana Krull

Institute of Linguistics, Stockholm University, S-106 91 Stockholm, Sweden

ABSTRACT

CV-sequencies occurring in the spontaneous speech of five Swedes were compared to the same sequencies reproduced by each speaker in citation form words. Two acoustic characteristics are reported here: F_2 trajectories for C= /b, d, m, n, l/ and voice onset time (VOT) for C=/p, t, k/. Plotting F_2 at the CV boundary (locus) as a function of F_2 within the vowel (nucleus) resulted in steeper regression lines for spontaneous 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 that the following vowel were considerably shorter in spontanaous 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[1].

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 pronounciation - "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. F_2, as a function of its frequency within the vowel results in a linear function called the "locus equation", Eq.(1):

$$F_{2i} = k * F_{2v} + c \qquad (1)$$

where F_{2i} denotes the initial locus of the second formant, F_{2v} 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 answer-

ing 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/ were located. F_2 was 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 F_2 locus and nucleus were nearer each other in spontaneous speech. However, further investigation showed that the locus and 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 I. y-intercept and slope for the regression lines of initial locus vs. nucleus F2 in CV-sequences.

Speaker	OE	RL	JS	PT	ÅV
C=/d,n,l/					
	n=142	n=83	n=107	n=118	n=88
Reference					
y-intercept	1106	1193	1041	1033	870
slope	0.29	0.31	0.36	0.36	0.45
Spontaneous					
y-intercept	937	936	823	795	755
slope	0.32	0.39	0.44	0.47	0.51
C=/b,m/					
	n=64	n=52	n=88	n=64	n=36
Reference					
y-intercept	568	313	363	276	389
slope	0.58	0.74	0.70	0.79	0.72
Spontaneous					
y-intercept	407	244	261	160	179
slope	0.64	0.75	0.75	0.83	0.81

2. VOT IN SPONTANEOUS SPEECH IND 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 position, utterance tempo, and vocalic environment. Stress, on the other hand, had a strong effect.

In Swedish, VOThas 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 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 where 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 in 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 dentals and labials. The mean VOT for the dental consonant was in most cases longer than that of the labial. For both

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.

	Closure	SD	VOT	SD	Vowel	SD	N
Speaker JS							
Unstressed CV							
spo	77	26	36	13	56	28	124
cit	163	47	34	12	135	41	
Stressed CV							
spo	63	17	46	13	79	37	28
cit	121	44	47	15	149	38	
Speaker PT							
Unstressed CV							
spont	90	34	35	12	56	30	101
cit	137	40	32	15	109	33	
Stressed CV							
spont	67	13	59	24	83	27	16
cit	72	15	56	9	129	34	

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 those 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 (Olle 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

[1] DUEZ, D. (1989). Second formant locus-nucleus patterns in spontaneous speech: some preliminary results on French. *PERILUS* X*,

[2] DUEZ, D. (1990). Some evidence on second formant locus-nucleus patterns in spontaneous speech in French. *PERILUS* XI*,.

[3] ENGSTRAND, O. (1983). Articulatory coordination in selected VCV utterances: A means - end view. (Ph.D. diss.). *Reports from Uppsala University Department of Linguistics* (*RUUL*) **10**, 76 ff.

[4] ENGSTRAND, O AND KRULL, D (1988). On the systematicity of phonetic variation in spontaneous speech. *PERILUS* VIII*, [5] Krull, D.

[5] KRULL, D. (1989). Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. *PERILUS* X*.

[6] KRULL, D. (1990). VOT in spontanous speech and in citation form words. *PERILUS* XI*,

[7] LINDBLOM, B. (1963). Spectrographic study of vowel reduction. *J. Acoust. Soc. Am.*, **35**, 1773-1781.

[8] LISKER, L. AND ABRAMSON, A. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, **10**, 1-28.

[9] LÖFQVIST, A. (1976). Closure duration and aspiration for Swedish stops. *Working Papers 13, Phonetics Laboratory, Lund University*, 1-39.

[10] MOON S.-J. AND LINDBLOM, B. (1989). Formant undershoot in clear and citation-form speech: A second progress report. *STL-QPSR* **1/1989**, Royal Institute of Technology, Stockholm.

[11] POCH OLIVÉ, D., FERNANDEZ-GUITIERREZ, N., AND MARTINEZ-DAUDEN, G. (1989). Some problems of coarticulation in CV stop syllables in Spanish and Catalan spontaneous speech. *Speech Research '89*, June 1-3, Budapest.

* Institute of Linguistics, Stockholm University

A STUDY OF [r] AND [ɾ] IN SPONTANEOUS SPEECH

Carme de la Mota Gorriz

Laboratori de Fonètica, Universitat Autònoma, Barcelona, Spain

ABSTRACT

This paper describes the acoustic features of the Spanish [ɾ] 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 [ɾ] and [r] has been the analysis of their duration, and therefore it is wellknown 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 (QUILIS [2]).This paper suggests some relevant differences in duration and spectral structure for both [ɾ] 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 militar service, etc.

These *data* have been compared with those obtained by studying [ɾ] and [r] in laboratory speech, that is, embeded in carrier sentences which were read at a normal speech rate by the same subject.

The registration was made in a sound proof room at the Phonetics Laboratory at the Universitat Autònoma de Barcelona, using a Revox A/77 tape recorder and a Shure 515 Sb Unidyne microphome.

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 [ɾ] and [r] have been studied in intervocalic contexts, either in stressed and unstressed syllables.

A whole of 300 items -uttered in labotatory 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-Vtransition starting point.

Intensity values are not studied in this paper. However, it should beinteresting to havealso into account the strong decrease in the sound pressure level of the consonant in futher research, as it has been pointed out by CHAFCOULOFF [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 [ɾ] and [r] are shorter in spontaneous speech, as is shown in Table 1.

TABLE 1: Mean duration values (in miliseconds). Comparison between laboratory and spontaneous speech.

<table>
<tr><td></td><td colspan="2">Laboratory Speech</td><td>Continuous Speech</td></tr>
<tr><td>[ ]</td><td colspan="2">31.6</td><td>27.1</td></tr>
<tr><td rowspan="4">[r]</td><td colspan="2">Total</td><td rowspan="4">45.31</td></tr>
<tr><td colspan="2">68.68</td></tr>
<tr><td>3 stages</td><td>5 stages</td></tr>
<tr><td>58.65</td><td>79.43</td></tr>
</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 significative 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 aproximately 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 [ɾ], there is a significative difference in miliseconds between spontaneous speech and laboratory speech. The mean values at Table 1 show the same differences observed at Figure 2 and Figure 3.

Both [ɾ] and [r] are shorter in continuous speech, although [r] is always the shortest one.

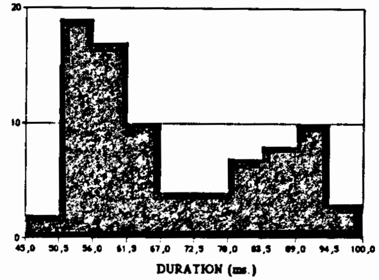


FIGURE 1: Histogram. Mean duration values for [r] in laboratory speech.

TABLE 2: Mean duration values (in miliseconds) for [r] in laboratory speech.

<table>
<tr><td rowspan="2">3 stages</td><td colspan="2">closed s.</td><td colspan="2">open s.</td><td>closed s.</td></tr>
<tr><td colspan="2">19.3
s.d.: 3.5</td><td colspan="2">18.2
s.d.: 2.2</td><td>18.1
s.d.: 3.7</td></tr>
<tr><td rowspan="2">5 stages</td><td>cl.s.</td><td>op.s.</td><td>cl.s.</td><td>op.s.</td><td>cl.s.</td></tr>
<tr><td>16.9
s.d.: 3.1</td><td>17.3
s.d.:2.5</td><td>18
s.d.:2.5</td><td>17.1
s.d.:2.4</td><td>16.8
s.d.:2.5</td></tr>
</table>

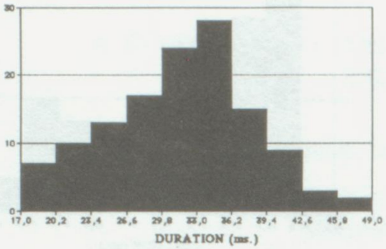


FIGURE 2: Histogram. Mean duration values for [ɾ] in laboratory speech.

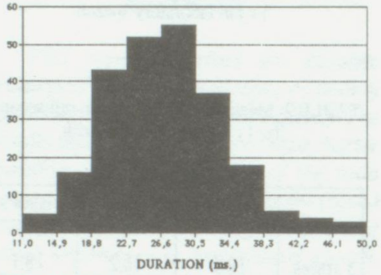


FIGURE 3: Histogram. Mean duration values for [ɾ] in spontaneous speech.

3. 2. Formant frequencies

The mean frequency values for the four first formants are those at Tables 3 and4. However, note that, as an hour of spontaneous speech reports us 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 [ɾ] in laboratory and spontaneous speech. (Hz.)

	Laboratory speech	Spontaneous speech
F4	3452.33	3409.11
F3	2304.06	2287.24
F2	1384.85	1354.29
F1	367.75	405.93

TABLE 4: Mean frequency values for [r] in laboratory and spontaneous speech. (Hz.)

	Laboratory speech	Spontaneous speech
F4	3397.05	3361.74
F3	2067.2	2025.49
F2	1129.48	1201.34
F1	434.4	443.13

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 5: Mean frequency values for [r] in laboratory speech (LS) and spontaneous speech (SS). Influence of the vocalic context on F1 and F2. (Hz).

F1	i/u	e/o	a
LS	313.48	383.31	456.83
SS	294.8	403	469.18
F2	o/u	a	e/i
LS	1023.6	1210.66	1703.84
SS	1038.86	1241.9	1600.4

TABLE 6: Mean frequency values for [r] in laboratory speech (LS) and spontaneous speech (SS). Influence of the vovalic context on F1 and F2. (Hz).

F1	i/u	e/o	a
LS	360.91	446.21	470.93
SS	no cases enough	428.15	496.66
F2	o/u	a	e/i
LS	1091.78	1089.93	1271
SS	no cases enough	1240.77	1396.27

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 taked 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 andthe following transition starting point. Comparison between spontaneous and laboratory speech.

	[ɾ]		[r]	
	e/i	o/u	e/i	o/u
LABORATORY SPEECH	10.7 %	2.9 %	7.2 %	1.9 %
SPONTANEOUS SPEECH	5.7 %	1.6 %	2.5 %	no cases enough

3.3. Spectral distribution

In fact, the results suggest that spontaneous speech favour the concentration of energy in the upper zones of frequency. About a third of the studied cases of [ɾ] in spontaneous speech show aperiodic energy in the higher frequencies, and about the ninety per cent of cases of [r] are periodic frictions. The fourth formant is the most intense in many cases. However, both [ɾ] and [r] are completely periodic in laboratory speech.

Figures 4 and 5 show some of the spectral differences observed between continuous and laboratory speech for [ɾ] and[r].

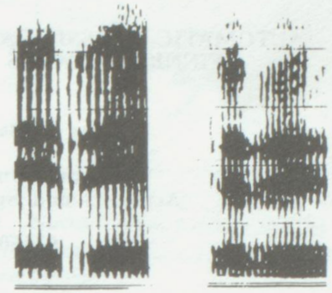


FIGURE 4: [ɾ] in laboratory speech and in continuous speech. Context: [e ɾ 'e].

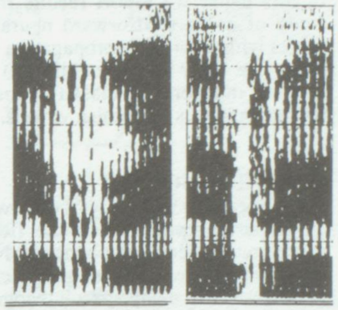


FIGURE 5: [r] in laboratory speech and in continuous speech. Context: [e r 'e].

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 thenext vowel. Further research should pay attention to intensity levels of [ɾ] and [r] in Spanish and to their spectral distribution in spontaneous speech.

5. REFERENCES

[1] FANT, G. (1960), *Acoustic theory of speech production*, The Hague: Mouton & Co.: 162-168.

[2] QUILIS, A. (1981), *Fonética acústica de la lengua españolaa*, Madrid: Gredos: 290-306.

[3] CHAFCOULOFF, M. (1979), "Les propiétés acoustiques de [j, y, w, l, r] en français", *Travaux de l'Institut de Phonetique d'Aix*, vol. 6: 10-24.

AUTOMATIC CLASSIFICATION AND FORMANT ANALYSIS OF FINNISH VOWELS 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 [1],[2]. They can be described as "black-boxes" that when 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 net topology where data flows from the input layer to the output layer via a hidden layer. Each layer is fully connected with the next one. The dimensionality of the net can be stated as the number of nodes in each layer (10-6-2 in figure 1).

This paper describes the application of neural networks to vowel recognition and analysis. Experimental results of vowel recognition and formant analysis are presented along with a summary regarding the usefulness of neural nets in this problem domain.

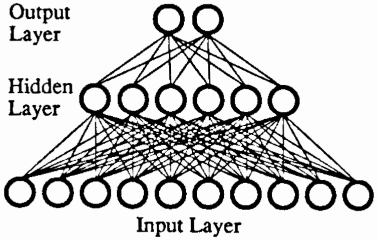


Figure 1. A possible network topology (general structure of a feed-forward net).

2. VOWEL RECOGNITION

For our vowel recognition experiments we used speech taken from 12 female and 24 male speakers. Static auditory spectra (288 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 /ä/. 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 measured 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 applied 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 /ä/ can also be seen in figure 2.

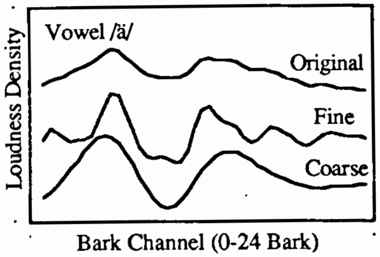


Figure 2. Original, fine, and coarse auditory spectrum representations of /ä/.

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.

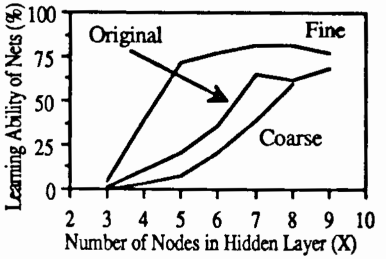


Figure 3. 48-X-8 Net's Ability to Learn 8 Vowel Spectra

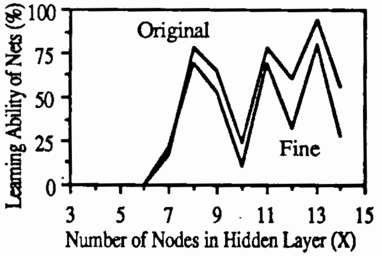


Figure 4. 48-X-8 net's ability to learn 192 male spectra.

2.1 Effect of F0 on Classification

A central part of the study was to see if the pitch frequency as additional information to the auditory spectrum could improve the classification performance of the nets by providing extra information for spectral normalization. For this test we created three training sets: male (24 speakers), female (12 speakers), and a male+female set (36 speakers), in order to see the degree of speaker independence and difficulty of the learning problem in each set.

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 F0 was 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

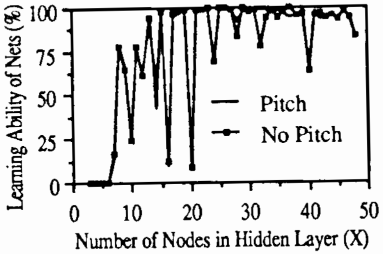


Figure 5. 24 male speakers with and without pitch information.

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, $X \in [2,15]$ 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 than 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 located 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.

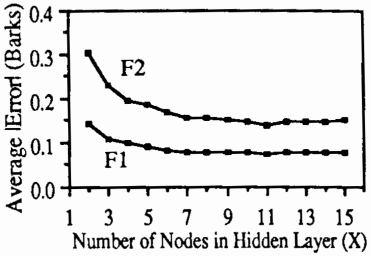


Figure 6. Average Formant Analysis Error of 64 Male Spectra as a Function of Net Size.

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 behaviour avoiding gross errors such as incorrect formant ordering, which is very difficult to achieve by traditional methods. We also observed that networks based the formant estimates on the general shape of the auditory spectrum but didn't 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.

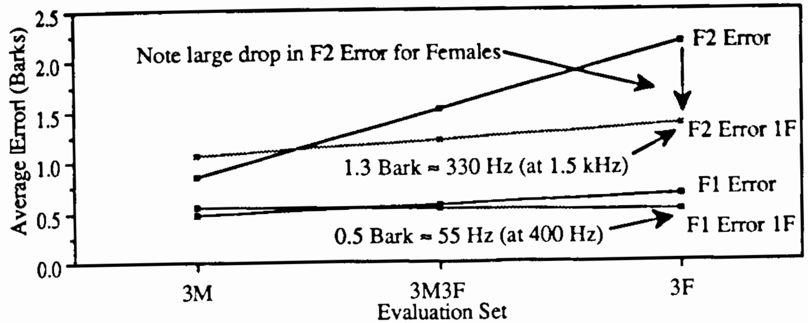


Figure 7. Evaluations of Trained Net on Independent Spectra.

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.

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 never 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 otherhand, 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

[1] LIPPMANN, R. (1987) "An Introduction to Computing with Neural Nets", In *IEEE ASSP Magazine*, 4.

[2] KARJALAINEN, M. (1987) "Auditory Models for Speech Processing", *11th ICPhS*, Tallinn.

[3] KARJALAINEN, M. et al. (1988) "QuickSig - An object-oriented signal processing environment", Proc. IEEE *Int. Conf. on Acoustics, Speech, and Signal Processing*.

[4] MUTHUSAMY, Y. et al. (1990) "Speaker-Independent Vowel Recognition: Spectrograms versus Cochleagrams", Proc. IEEE *Int. Conf. on Acoustics, Speech, and Signal Processing*.

[5] WAIBEL, A. et al. (1989) "Phoneme Recognition Using Time-Delay Neural Networks", In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3).

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 Bulgarian 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 together with their cluster statistics: means, standard deviations, maximums, minimums, skewnesses and kurtosises.

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 (b-VOWEL-b) allophones of the vowels: /i/, /e/, /ə/, /a/, /u/, /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 uncorrect utterances to be rejected by the computer and only the correct ones to be admited 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 performes a FORTRAN program based on the subroutine FORIT from the SSP [4] and builded up according to the algorithm [1]. The computer produces, except of output listings of the labeled points

(FO, F1, F2, F3) ..../*/
LABEL

where

LABEL = /phonemic 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. 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 uncorrect 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 /ˋi/ cluster above the upper limit of the graph and one more in the cluster of the vowel /ˋa/. The number of coinciding and out-of-the-graph points is presented in the last column of Tab. Tab. 1 to 4.

3. CLUSTER STATISTICS

The statistical processing of all vowel utterances verified by the listeners is performed by a FORTAN program which makes extensive use of the SSP subroutines [4], among them the subroutine TALLY to compute means, standard deviations, maximums and minimums and the subroutine MOMEN to help by the computing of the skewnesses and curtosises. These statistical estimates, computed for each cluster of equally labeled points, are presented in Tab. 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 comunity by trivial and computerized comparative phonetics studies and by machine synthesis and recognition of Bulgarian speech.

6. REFERENCES

[1] CHRISTOV, Ph. (1983), "An algorithm using linguistic information and Its application to the analysis of speech in the spectral domain", Proc. XI ICA, Paris, 4, 161-164.

[2] CHRISTOV, Ph. (1987), "Computer aided analysis of stressed and unstressed Bulgarian vowels from 30 male and 30 female speakers", Proc. XI ICPhS, Tallin, 3, 121-124.

[3] CHRISTOV, Ph., (1987), "A large Bulgarian central allophones data base", Proc. XI ICPhS, Tallin, 5, 232-235.

[4] IBM (1970), Application program system/360 Scientific subroutine package, Version 3-th, Programmer's manual, 5-th edit., IBM Techn. Publ. Dpt., New York.

APPENDIX:

Word list in rough phonemic IPA - transcription:

STRESSED	UNSTRESSED
/bˋiblija/	/biblˋejski/
/bˋebe/	/bebˋet ef/
/bˋaba/	/babalˋək/
/bˋəbrek/	/bəbrekovˋiden/
/bˋobof/	/bobˋovina/
/bˋuba/	/bubˋar/

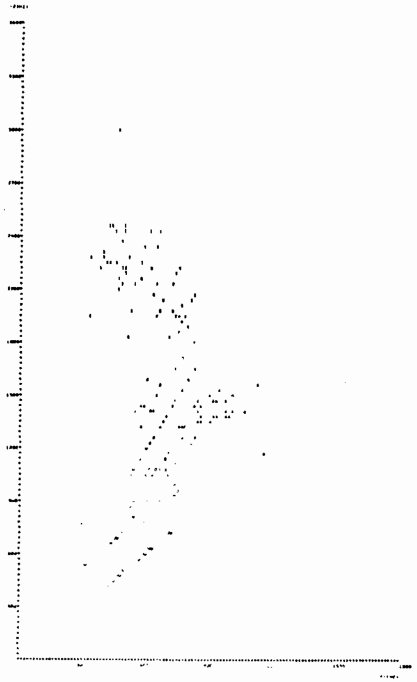


Fig. 1. First two formants computer graph of the six Bulgarian vowels uttered in stressed position by 30 male speakers

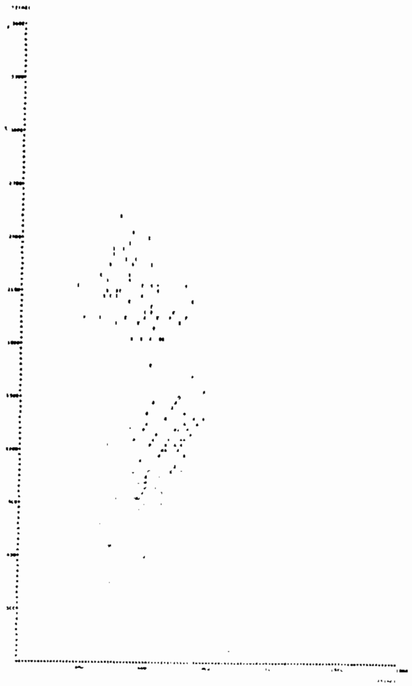


Fig. 2. First two formants computer graph of the six Bulgarian vowels uttered in unstressed position by 30 male speakers

TERMINOLOGY:

CLUSTER - a group of objects put together by some resemblant feature (DURAN, B., ODDEL, P. (1974),"Cluster analysis. A survey", Springer Verlag). The term is familiar in the theory of pattern recognition.

LEGEND TO THE FIGURES:

In the computer print outs 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/ :
@ = /ə/ :	O = /o/ :

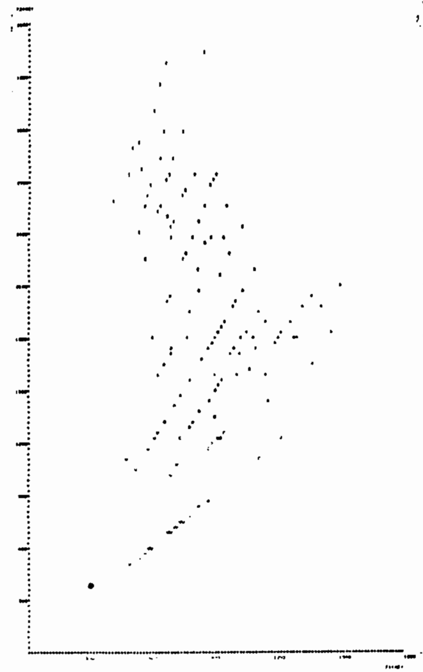


Fig. 3. First two formants computer graph of the six Bulgarian vowels uttered in stressed position by 30 female speakers. There are three points /*/ with rather high second formant which got aut of this graph. Two of them are labeled as /`i/, (F1=624, F2=3744) and (F1=850, F2=3825), and one as /`a/, (F1=1176, F2=3864)

REMARK:

With a single exception (See text to Fig. 4) coinciding points belong to one and the same vowel cluster.

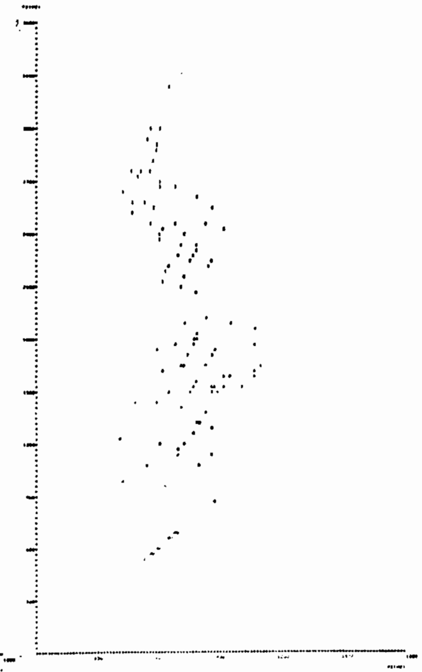


Fig. 4. First two formants computer graph of the six Bulgarian vowels uttered in unstressed position by 30 female speakers. The point /ə/, (F1=768, F2=1536), which coincides with a point of the /a/-cluster, is not marked on the figure

LEGEND TO THE TABLES:

n - number of vowel utterances admited 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

PHONOLOGY OF SYNHARMONISM AND A NEW SYNHARMONIC SCRIPT

A. Dzhunisbekov

Institute of Linguistics
Academy of Sciences of the Kazakh SSR

Turkic phonology is the phonology of synharmonism. The model of the phonology of synharmonism 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 orthoepic 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 inturkic 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 Turkic languages. It is a specific language unit forming the integrity of syllables and words in turkic speech.

Here is the model of synharmonism, built as a "circle" because both synharmotypes (palatal and labial) as well as all the four synharmotimbres (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 synharmotimbres 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 synharmonism. 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 synharmotimbres in the system of consonantism. Such universality of consonants (in contrast to vowels) permits to use them as basic sounds in constructing the synharmonic script.

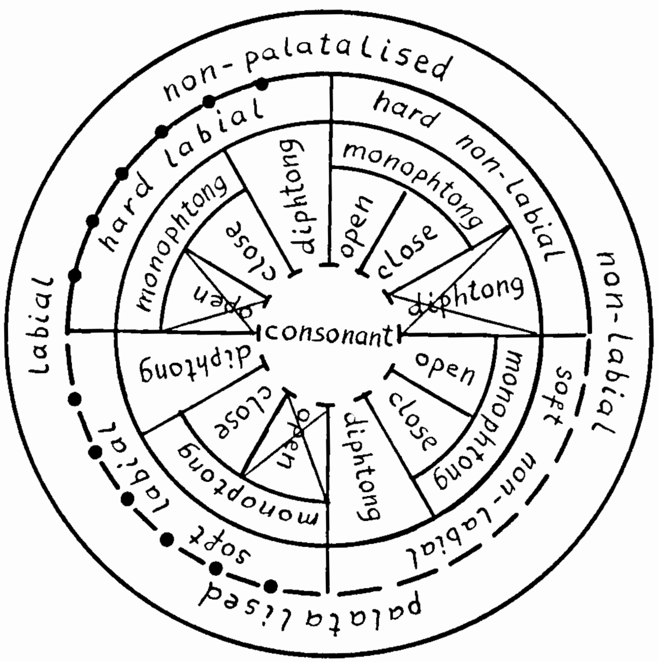
The level of formalism of the proposed model may be subjected to criticism, and we shall be glad if someone will manage to give more efficient and accurate definition of synharmonis and to construct the appropriate variant of the model. We want, to remind that nobody succeeded in constructing a good working model of synharmonism 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 "europocentristic" 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 synharmonism turned out to be insufficient and erroneous.

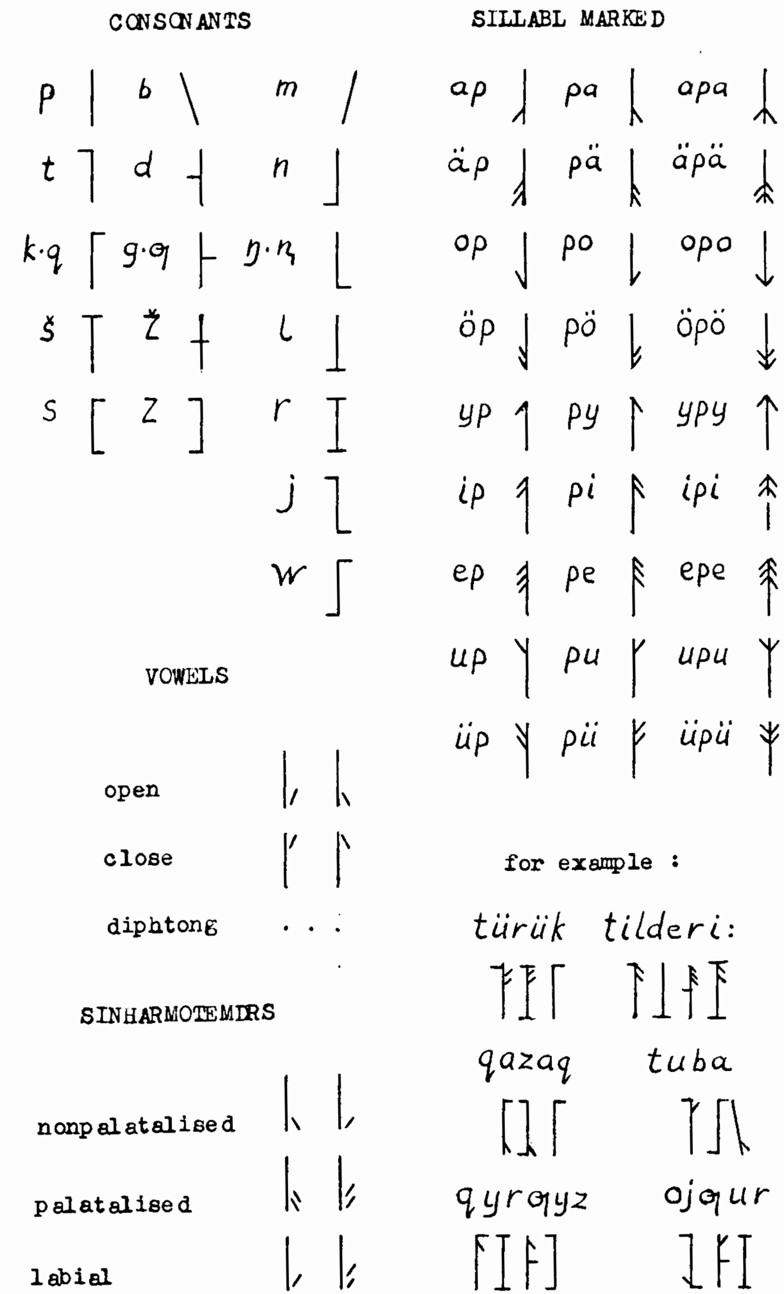
So far as synharmonism is the phonological basis of the proposed system of the script, we use the simplified term "syngramma" for designating the syllabosymbol. Graphs of syngrammas are elementary: they consist of the joining of only straight lines (we intentionally avoided round, oval, curved and other complex lines for the scripts) and there are only three of them. Each syngramma consists of the combination of the three straight lines: vertical line " | " which is basic for all syngrammas; horisontal " – "-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





DE L'INDÉPENDANCE DU PHONÈME FAIBLE AU SYSTÈME PHONOLOGIQUE DE LA LANGUE RUSSE

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 phoneme in the phonological systeme 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 phonètique 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. 1956 a. [2] à savoir à la thèse de la connexion de la phonologie avec la morphologie, du phonème avec la morphème de la série de phonèmes avec l'indentité 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 décrivant 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-a-dire n'embrasse pas la majorité des cas. Le plus souvent le phonème faible est présentée 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 conformement 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 forts et 2 phonèmes voyelles faibles /ʌ/ et /ʌ$_1$/.

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-a-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 mous. Par ex., putь, phonol./put$'_2$/, /t'_2/ - chez Avanesov R.I. est le phonème faible de sourdité-sonorité mais fort de dureté-molesse et chez nous - /t'_2/ - 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 púty où le phonème faible de sourdité-sonorité /t_2/ 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 présentée comme le phonème faible de tel ou tel signe et comme le phonème fort d'autre signe. Par ex., nous concidérons la position de la distinction maximale pour les phonèmes forts comme la position précédant les voyelles excepté /e/ à la limite du thème de la flexion où sont présentés les phonèmes consonnes faibles de dureté-molesse (dans cette position la distinction des consonnes de la dureté et de la molesse est absente). Devant les voyelles tous les phonèmes consonnes forts sont opposés (par ex.,/p/ar- /b/ar, /f/ar de fara,/p/or de pora et porá - /b/or- /m/or - /v/or - /s/or 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 consonnes) se présentent les phonèmes forts (surtout non-appariés de tel ou tel signe) ainsi que (principalement) les phonèmes consonnes faibles. Nous distinquons la distribution nette dans les positions: devant /e/ à la limite du thème et de la flexion - les phonèmes faibles de dureté-molesse (par ex., /na/ sto/l_1é/, /o/ so/f_1é/, /na/ ko/r_1é/), à la fin du mot - non-apparié /c/, /č'/ /h/, les sonores dures-molles et /j/ et les phonèmes consonnes faibles de sourdité-sonorité; devant les consonnes se présentent surtout les phonèmes faibles (de dureté-molesse, de sourdité-sonorité et des phonèmes faibles de deux signes) et aussi les phonèmes consonnes forts non-apparies et devant /m/, outre cela - /t/, /t'/, /d/, /d'/, /s/, /s'/, /z/, /z'/.

1.5. La reconnaissance du phonème faible en qualité de l'unité phonétique indépendante, et la renoncement a l'emploi de la série de phonèmes entraîne le renoncement de la transcription morphophonématique [2, §77, p.221-224]. Nous proposons d'employer seulement la transcription phonologique c'est-à-dire la transcription qui présente l'aspect phonèmatique de la forme du mot. Elle présente aussi la composition phonèmatique du morphème 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 phonème faible hors de la série des phonèmes permettent d'étudier du point de vue de la combinaison et du rendement fonctionnel une grande couche de lexique russe où sont présentés tous les phonèmes faibles vérifiés ou non par la position forte. On peut donner le tableau complet des possibilities combinatoires et fonctionnelles. Les phonèmes faibles aussi que les phonèmes forts sont le composant indépendant des formes du mot. Par ex., /p/ol, /s_3/tol, /z_1/r'a, /s_1/razu etc. Ils possédent comme les phonèmes forts la fonction de distinction sémantique et le rendement fonctionnel 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 morphèmatique. Par ex., les phonèmes consonnes faibles de dureté-molesse /p_1/, /b_1/, /t_1/, /d_1/, /s_1/, /k_1/, /g_1/ possédent le rendement fonctionnel, les autres dans cette position, ont le rendement fonctionnel relâché; dans la position à la limite du thème et de la flexion tous les phonèmes faibles de dureté-molesse ont le rendement fonctionnel relâché. Les phonèmes consonnes faibles de surdité-sonorité /f'_2/, /$š'_2$/, /k'_2/ possédent aussi le rendement fonctionnel relâché dans cette position. Tous les phonèmes consonnes faibles de deux signes ont le rendement fonctionnel relâché dans la position à l'intérieur du morphème. Ainsi la reconnaissance du phonème faible en qualité de l'unité phonétique indépendante permet de construire le modèle phonologique où les phonèmes forts et faibles seront présentés comme unités phonétiques indépendantes.

REFERENCES

[I] АВАНЕСОВ Р.И., СИДОРОВ В.Н. (I945), "Очерк грамматики русского литературного языка", М., ч.I, с.39-67.
[2] АВАНЕСОВ Р.И. (I956). "Фонетика современного литературного языка", М.
[3] ДМИТРЕНКО С.Н. (I988). "Фонемы русского языка. Их сочетаемость и функциональная нагрузка". Автореф. докт. дисс. М.

MODELLING VOWEL SYSTEMS BY EFFORT AND CONTRAST

L.F.M. ten Bosch

Institute of Phonetic Sciences, University of Amsterdam, The Netherlands

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. It results from accurate regulation of the subglottal air pressure and, at the same time, manipulation of the glottal and vocal tract muscles. Phonemes such as vowels and consonants act as linguistic (phonological) units in a language, but at the same time, the corresponding allophones are 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 manner in which 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 vowel models.

2. INTERNAL STRUCTURE

Vowels in language principally serve a linguistic goal. Their existence helps to distinguish words semantically, which is clear in the case of minimal word pairs. Historical linguistics 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 Germanic languages), to whims of fashion (some cases of diphthongization), to ease of articulation (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 modelled 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 (Disner, 1983).

In vowel models, the actual linguistic vowel systems are assumed to optimize perceptual contrast and, in an extension, articulatory ease. Liljencrants & Lindblom (1972) were the first to implement a principle of optimal perceptual contrast in a so-called vowel dispersion model. In a 2D formant space, vowels were positioned such that the system contrast was maximized, by the minimization of the system quality Q:

$$Q = \sum \frac{1}{d(v_i, v_j)^2} \qquad (1)$$

where $d(v_i, v_j)$ denotes the (Euclidean) distance between any two vowels v_i and v_j in the 'perceptual space', and the sum is taken over all distinct vowel pairs ($1 \leq i < j \leq N$). The particular implementation chosen by L&L suffered from the drawback to generate too many high vowels for large N, due to a too large perceptual distance between /i/ and /u/.

One of the basic ingredients in this approach, viz. the perceptual distance between two vowels, has later been modified to more sophisticated submodels for the auditory spectrum (Bladon & 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 dependency 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 bark transformation of F_1 and F_2 (Kewley-Port & Atal, 1989). Since their stimuli, however, were determined by two parameters only, this result must be carefully interpreted, leaving aside the question about the relation between the phonemic distance (that we search) and the phonetic distance (that they measure).

While d has been subject to continuous refinement, the *system contrast* Q, however, has not grown beyond the form

$$Q = \sum \frac{1}{d^2} \qquad (2)$$

d now involving combinations of transformed formant frequencies (Schwartz *et al*, 1989) or spectral differences (Lindblom, 1986)). The problem that we want to address here is that this expression Q is in fact very arbitrary, it being suggested by repelling forces between magnetic monopoles or dipoles, but lacking, in fact, any linguistic or even physical basis. Ten Bosch *et al.* (1987) propose an expression

$$Q = \prod (1 - \exp(-\alpha d)) \qquad (3)$$

the product being taken over all distinct vowel pairs, and α some scaling parameter. Q is to be optimized. The rationale is, that the factor $1 - \exp(-\alpha d)$ ($\equiv \pi(d)$) is interpretable as a probability of two vowels on a distance d not being confused. The system quality Q would then 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. Also in this approach, however, a weak argument can be detected, namely that the resulting optimal vowel configurations can (easily) be shown to be dependent on the exact shape of $\pi(d)$ (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:

$$Q = \min_{i \neq j}\{d(v_i, v_j)\} \qquad (4)$$

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' in 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 is done in eq. 2). (c) Any sufficiency constraint of the system contrast is directly relatable to the minimal perceptual distance between vowels. The systems, obtained by optimizing eq. 4, are similar (but not equivalent) to the ones, obtained by minimizing eq. 2 (Ten Bosch, 1991). This yields, in my opinion, a strong argument for the latest modified Q (Ockham). Property (a) is particularly useful in numerical simulation of push and drag chains. In 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 diphthongeal properties in detail, gross effects, such as the preference for diphthongs to have a relatively large trajectory, can be clearly demonstrated.

Articulatory constraints were not explicitly dealt with in these models: all calculations were carried out in the acoustic domain. Recent implementations attempted to combine perceptual and articulatory constraints (Bonder, 1986; Ten Bosch, Bonder & Pols, 1987; Ten Bosch, 1991). Other approaches were carried out by Abry, Schwartz, Badin, Boë, Perrier, Guérin (see the references) and colleagues in Grenoble. Stevens (1989) has put 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, other points of view have been adopted (leading to e.g. the notion of focal points, articulatory plateaus, sufficiency instead of optimality), and more elaborated definitions of Q have been introduced (Schwartz *et al.*, 1989).

Ten Bosch *et al.* (1987) propose a vowel system model based on maximal acoustic contrast together with a minimal articulatory effort criterion, by minimizing

$$D_A^2 + S \cdot (Q-1)^2$$

where D_A is the total articulatory system effort, Q given by eq. 3, and S a slack variable as used in optimization problems (S being a large positive number). This 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 here leave aside the definition of 'articulatory system effort' and even forget the role of consonantal context in any definition of articulatory ease.
Another important goal is the refinement of the overall articulation-to-acoustics relation. The Quantal Theory (QT; Stevens, 1972, 1989) makes use of the non-uniformity of this 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 likely be a member of a vowel system. The presuppositions 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 convincing (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-acoustics mapping and probably, the categorical perception of speech sounds. If quantality exists, it is probably a result of close approximations of formant frequencies (Stevens, 1989; Badin *et al.*, 1990; Schwartz *et al.*, 1989; Ladefoged *et al.*, 1988).

We briefly return to the open question of phonological enrichment of phonetic vowel models. An unsolved, and perhaps unsolvable, drawback inherent to phonetic models is that they cannot easily account for the linguistic demand for vowel contrast, although linguistic oppositions are ultimately based upon phonetic contrast. Is there a relation between the need of intervowel contrast and the 'lexical load' of the opposition? The relation between phonetic contrast and phonological contrast seems not to be derivable directly from the statistics on lexemes in a language. In Dutch, /ɑ/ and /ɔ/ have the largest (most often frequented) minimal set in common, despite they are a very close pair in the Dutch vowel system.

3. EXTERNAL STRUCTURE

We mean by external structure of vowel systems the description of the vowel space boundary in articulatory terms. It opposes the internal structure, with which we mean the positional organization of the vowels themselves. External structure is related to the notion of 'possible speech sound' (Lindblom, 1990). From a phonological point of view, the boundary of the vowel space is globally anchored between the combinations [low], [back, round] and [front, unround], representing /a/, /u/ and /i/, respectively. From a phonetic point of view, the set of possible speech sounds is a subset of the total sound-producing potential of the vocal tract. The relation articulation-to-acoustics and the inverse problem, the computation of the vocal tract shape from the acoustic output, plays here a central role.
The problem, how to relate vocal tract shape and acoustic output can be tackled in different ways: (1) in terms of electric LC-circuits. Historically, this has been the usual paradigm; (2) in terms of the n-tube representation of the tract (Fant, 1960; Atal *et al.*, 1978; Bonder, 1983; Ten Bosch *et al.*, 1987; Stevens, 1972, 1989). (3) in terms of articulatory-based tract models (by Lindblom, Sundberg, Ladefoged, Mermelstein, Maeda). (4) in terms of eigenfunctions of the Webster horn equation (Karal, 1953; Mrayati *et al.*, 1988).
Apart from their starting points, these four approaches are in fact mathematically equivalent.

Perrier *et al.* (1985), using Maeda's statistical analyses of articulatory positions has shown that the boundary of the vowel triangle can adequately be simulated by putting specific lower and upper bounds on the tube segment areas. Bonder (1983) and Ten Bosch (1991) studied this phenomenon by using the n-tube as articulatory model.

Since Atal *et al.* (1978), it is well known that the inverse problem has no unique solution (fibre). In order to specify one unique exemplar from the fibre, additional constraints have to be defined. This provides us the possibility to define an effort value to each formant position. The acoustic output being given, let ϕ denote the corresponding fiber of all positions x in the articulatory space. Furthermore, we have some articulatory effort function e defined on the articulatory space. Then

$$\min\{e(x) \mid x \text{ on } \phi\}$$

denotes the minimal effort value on the fiber. This value depends on the fiber, i.e.

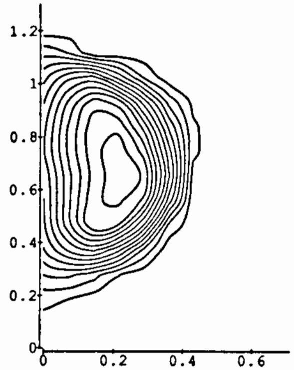


Fig. 1. Contour lines in the $(x, y) = (F_1, F_2)$ plane of an effort function defined on the articulatory space. Scaling: $1 \equiv 2000$ Hz.

the acoustic output. Accordingly, the minimum effort value defines a 'effort landscape' on the acoustic space. It is shown in Ten Bosch (1991) that a relatively simple effort function e 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

ATAL B.S., Chang, J.J., Mathews, M.V., and Tukey, J.W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J. Acoust. Soc. Am. 63. p. 1535 - 1555.
BADIN, P., Perrier, P., Boë, L.-J., and Abry, C. (1990). Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. J. Acoust. Soc. Am. 87. p. 1290-1300.
BLADON, R.A.W. & Lindblom, B. (1982). Modeling the judgment of vowel quality differences. J. Acoust. Soc. Am. 69. pp. 1414-1422.
BONDER, L.J. (1983). The n-tube formula and some of its consequences. Acustica 52, p. 216 - 226.
BONDER, L.J. (1986). A prediction method for modal n-vowel systems. Procs. Inst. Phon. Scs. Amsterdam, vol. 10. p. 73-90.
TEN BOSCH, L.F.M., Bonder, L.J., and Pols, L.C.W. (1987). Static and dynamic structure of vowel systems. Procs. 11th Intern. Congress Phon. Scs., vol. 1. p. 35-238.
TEN BOSCH, L.F.M. and Pols, L.C.W. 1989). On the necessity of quantal assumptions. Questions to the Quantal Theory. Journal of Phonetics 17. p. 63 - 70.
TEN BOSCH, L.F.M. (1991). On the structure of vowel systems. An extended dispersion model. PhD-thesis (in preparation). University of Amsterdam, The Netherlands.
DISNER, S.F. (1983). Vowel quality. The relation between universals and language-specific factors. UCLA WPP 58. Un. of California, LA.
FANT, G. (1960). Acoustic Theory of Speech Production. Mouton & Co., 's-Gravenhage.
KARAL, F.C. (1953). The analogous acoustical impedance for discontinuities and constrictions of circular cross section. J. Acoust. Soc. Am. 25. p. 327 - 334.
KEWLEY-PORT, D., and Atal, B. (1989). Perceptual differences between vowels located in a limited phonetic space. J. Acoust. Soc. Am. 85. p. 1726-1740.
LADEFOGED, P. and Lindau, M. (1988). Modeling articulatory-acoustic relations. UCLA Working Papers 70. p. 32 - 40.
LILJENCRANTS, J. and Lindblom, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. Language 48, p. 839 - 862.
LINDBLOM, B. (1986). Phonetic universals in vowel systems. In: Experimental Phonology (J. Ohala and J. Jaeger, eds.). Academic Press, Orlando, Florida. p. 13 - 44.
LINDBLOM, B. (1990). On the notion of "possible speech sound". Journal of Phonetics 18. p. 135 - 152.
MRAYATI, M., Carré, R., and Guérin, B. (1988). Distinctive regions and modes: A new theory of speech production. Speech Communication 7. p. 257 - 286.
PERRIER, P., Boë, L.J., Majid, R., and Guérin, B. (1985). Modélisation articulatoire du conduit vocal: exploration et exploitation. In: Proceedings of the 14èmes Journées d'Etudes sur la Parole (Groupement des Acousticiens de Langue Française, Paris). p. 55 - 58.
SCHWARTZ, J.L., Boë, L.J., Perrier, P., Guérin, and Escudier, P. (1989). Perceptual contrast and stability in vowel systems: a 3-D simulation study. Proceedings of the Eurospeech Conference, Paris. p. 63-66.
STEVENS, K.N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data. In: Human communication: A unified view (E. David and P. Denes, eds.). McGraw-Hill, New York. p. 51 - 66.
STEVENS, K.N. (1989). On the quantal nature of speech. Journal of Phonetics 17. p. 3 - 45.

ARTICULATORY AND PERCEPTIVE ASPECTS OF TYPOLOGY OF SOUND SYSTEMS IN CONDITIONS OF MULTILINGUISM

T.Verbitskaya and T.Korolyova

State University, Odessa,USSR

The problem of interaction between perceptive and articulatory aspects in foreign language teaching 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 typologycal identity is suggested. The experimental results show the effectivity of the method in teaching production and perception of foreign language under the conditions of bilinguism.

Lately linguistics treats the problem of language systems interaction under conditions of bilinguism and multilinguism quite often. One of the most important aspects in this respect is the way the languages are to be studied. There is a natural bilinguism that takes place under conditions of direct communication between the speakers whose mother tongue is not the same, and artificial bilinguism which occurs under particular circumstances when a foreign 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 to the accentless mastering of the language. In this situation the language systems interaction is "reduced" to influence of the pupil's mother tongue (MT) on the foreign language (FL) studied. In case of natural bilinguism a mutual effect of two contacting languages can be noticed. Both forms are to be closely examined by linguists and specialists of adjacent sciences. Natural multilinguism is characteristic for the situation which has taken place in the Soviet Union. When Russian and national languages interact, their influence on each other is doubtless: 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 realisation are caused by the Russian language used in the Republic but not the national language of the Republic itself. It is known that the artificial form of contacts can be regarded as a satisfactory model of natural processes and the other way about - many advantages of natural contacts can be used to facilitate teaching process. The analysis of interference problems in linguistic aspect requires to list the potentially expected forms that can appear under the circumstances. This can 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 main theoretical and 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 points of view on the problem of interrelation between perceptive and articulatory aspects lies in the question of double interference as a source of accent under the conditions of mixed bilinguism because phonology of hearing conditions the phonology of speaking. In a number of works perceptive aspect is looked upon as a basical one.Thus perceptive basis (PB) occupies a particular place alongside with the articulatory one (AB); PB is a unity of memory stored patterned phonetic units and the rules of comparison with them.

Distinction in various languages can be explained by difference in patterns of phonemes, patterns of supersegmental units and the rules of comparison with the patterns.In accordance with modern concept of mechanism of speech perception phoneme is of primary importance in making up one's mind. Trubetskoy N.S.[4.59] underlined that the perception of foreign language sounds is phonologically conditioned. It does not mean that a foreign language speaker interprets any unfamiliar sound as a known one, i.e. turns any sequence into a sequence of native language phonemes. On the basis of the experimental data available we can state that a man is capable of differentiating a larger number of sounds than the amount of phonemes in his MT. Nevertheless, this capability of a person is also conditioned by phonological relations. When perceptiong the sounds which are absent in the MT the examinee does not always place them as phonemes belonging to his MT: rather a subtle differenciation is possible. It grounds on the properties of perceptual processing of the sound signals, knowledge of one or several of FL and individual capabilities of the examinees as well.

Everything mentioned above testifies to the necessity to specify the traditional view on the perceptive abilities of a man. In this connection one of the important tasks is to find out the phonetic characteristics of sound sequence which are used in the act of perception. At the same time it would be wrong to forget that perception and (re)production are the two sides of joint activity. Numerous attempts to reveal the dependance of gained results on the acoustic characteristics of perceived sounds testifies against presence of direct connection between perceptive & acoustic correlates. The experimental data illustrate the presence of tight correlation between perception and articulation [1]. There is much interest in the results of the experimentson speech sounds perception which allowed to come to the conclusion that perceptive system is characterized by specific dependance on certain languages,on the one hand,and common universal traits, on the other hand. For example, it is believed that the universality of perceptive system manifests itself in the fact that the most "hurtful" in languages are the "prominent" vowels /i-a-u/ et al. However, as the experimental data evidences the relation of specific and universal is beyond the limits of voice tone acoustic activity (in particular with respect to the sounds mentioned). The role of mechanism of contacting languages interaction is much more important. If the main postulate of this research reads that deviations from the normative pronunciation of

the language studied in the speech of a bilingual (polilingual) results from the contact of phonetic systems in MT and FL. Such "hurtful" points can be ticked out on the basis of mechanisms of their interaction in contacting languages. Thus, there are attempts to group phonemes in MT and FL on the principles of their acoustic or articulatory similarity; one can find the following defenitions:"phonemes of close group","phonemes of relatively close group", "phonemes of distant 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 heterosystemic languages, and the conclusion about coincidence of the whole number of English sounds /p/,/b/,/f/,/s/ with the corresponding Russian phonemes is not convincing, because two things are ignored: a clearly distinguishable opposition in the system of Russian consonants-palatal/hard sounds which do not occur in English (German) language and phoneme distribution. At the same time the correlation of paired palatal-hard phonemes plays the leading role in Russian consonatism because, firstly, the largest amount of consonant phonemes are involved in this opposition and, secondly, palatal-hard consonants influence greatly the adjacent vowels and cause the whole range of peculiarities in allophones of vowel phonemes in speech.
Now it is just to the point to say that when carrying out a typological research not only the existance of this or that fact is to be taken to consideration but the place it ocupies in the system of the language under discussion.
Taking the above-mentioned fact into account we can't consider such phonemes as German /s/ or /f/ and Russian /s/,/s'/ and /f/,/f'/ as identical ones, though there is some definite articulatory similiarity between them. The difficulty of mastering the identical German consonants is caused by the following peculiarities of Russian articulatory basis: 1) in position f+ +nonfront vowel or voiceless fricative consonant there is no additional raise of the back part of the tongue; 2) in position f+front vowel there is no additional raise of the middle of the tongue; 3) in position f+voiced fricative there is no regressive assimilation.
The latter case causes much difficulty for reproduction because of presence of two different processes in the systems of Russian, Ukrainian consonatism, on the one hand, and German consonatism, on the other hand: regressive assimilation in voicing and progressive assimilation in devoicing, compare A(vg)anistan-A(fg)anistan. It would be a mistake to come to the conclusion that there are no difficulties in reception of the sounds alike. The results of perceptive tests show that the opposition voicing-devoicing in the German language is not so strong as in Russian or Ukrainian [5.132], that is why the differentiation of German voiceless /Fortes/& voiced /Lenes/ consonants gives rise to many mistakes even in intervocal position (e.g. reise-reiße).
It is a good thing to mention here A.Reformatsky's words: "The only place where phonology does not feel at home is the methods of teaching pronunciation of a foreign language. Quite often separate sounds as well as imaginary identical sounds are of special attention here" [2.506].
One of the key problems is to define distinct criteria of typological identity in groups of sounds in the languages being compared. The degree of similiarity (difference) of the elements in articulation of the compared sounds can be regarded as such a criterion. It allows to group the sounds accurately enough according to the degree of their typological likeness.
What is the degree the articulation elements should differ to place the sound of a FL in the group which has no analogue in MT? A clear-cut trait, which shows the absence of analogue, is incompatibility of articulatory actions in the compared languages. Thus, in the field of vocalism in Russian, Ukrainian and English articulatiry basses the following articulatory actions are incompatible: labialization+forelingual articulation. This situation excludes the possibility of labialized forelingual vowel appearance. in the phonological system of the languages mentioned above in contrast to the systems of German and French vocalism where these types of vowels are available. In Russian & Ukrainian articulatory bases consonatism is characterized by incompatibility of the velum moving down & backlingual articulation which takes place when pronouncing English and German backlingual consonant [ŋ].
The opinion that the sounds which have no analogues in MT are mastered worse - is widely spread nowadays. Nevertheless, Scherba L.V. warned that particular difficulties are connected with the sounds which have analogues in MT. The sounds that have no analogues in MT attract our attention & thus are not identified with either sound in the MT and do not cause ambiguity [3].

The Number of Perceptive and Articulatory Mistakes at Different Stages of Students Linguistic Competence, %

Group No	Sound	Before phonetical course: perception	Before phonetical course: articulation	After phonetical course: perception	After phonetical course: articulation	After a five-year course: perception	After a five-year course: articulation
1	1	21	38	6	5	8	10
	g	28	42	5	6	8	13
2	ч	8	48	3	4	1	3
	ŋ	7	55	3	2	0	1

The results of tests carried out give us grounds to admit this statement to be true since the sounds which have no analogues in MT are differentiated better than the ones having certain correspondence in MT. It can be explained by the listener's attempt "to fit" the perceived signal into the zone of standard sound-type in MT identical as to its acoustic characteristics. When practicing pronunciation the similiarity of MT and FL is imaginary and "provocative" [2.510]. This regularity is illustrated by the data given in the table.
It is demonstrated that German sounds /i/,/g/ which have relative analogues in Russian and Ukrainian languages are worse mastered in perceptive and articulatory aspects than sounds /ч/ and /ŋ/ which have no analogues. There is a clearly seen tendency within the sounds of the first group to increase the amount of errors in sound perception and articulation, particularly, when purposeful training is over. The skills of oral perception and production of sounds in the second group are much more stable.
The use of the criterion of compatibility/incompatibility in articulatory actions at the process of sound production allows to reveal a clear-cut correlation between articulatory and perceptive aspects of speech under multilingual conditions.

[1] LIEBERMAN, A.M.(1967), "Perception of Speech Code", Psychological Revue,75-86.
[2] REFORMATSKY,A.A.(1970), "Phonology at service of teaching foreign language pronunciation",From the history of national phonology, Moscow,506-511.
[3] SCHERBA,L.V.(1957),"Phonetics of the French language",Moscow.
[4] TRUBETSKOY,N.S.(1960), "The basis of phonology", Moscow:Izd.Inostr. Lit.
[5] WIEDE,E.(1981),"Phonologie und Artikulation sweise in Russischen und Deutschen" Leipzig:VEB Verlag Enzyklopadie.

THE INFLUENCE OF SOCIAL FACTORS ON URBAN SPEECH

L.V.Ignatkina

Leningrad State University, USSR

ABSTRACT

The influence of social factory (education and profession) on urban speech in twenty cities of Russia is discussed.

One of the important research area in modern Linguistics is the study of the standard variant of a national language and the factors which influence modification of a speech sound.

This problem is closely connected whith the study of the development of standard pronunciation, geografical variability and social factors of language 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 macrosociolinguistics, using P. Bell's definition and it is a part of sociolinguistic research fulfilled in the USSR, Poland, Chechoslovacia, Germany and the USA by the linguists of different countries. The main attention is devoted to the factors of edu cation 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- and South-Russian, of 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-examin the functioning and the development of the orphoepic 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 proccess of contacting between standard language and other forms of national language (local dialects, popular speech).

The towns and cities observed are situated within (territories of dissemination) of the dialects which have different character and different time of origin. Archangelsk and Vologda are within the zone of functioning of the Northern Russian dialects.

Krasnodar, Kursk, Rostov-on-Don, Ryazan, Simfero pol - South-Russian dialects, Volgograd, Nizhny Novgorod (Gorky), Samara (Kuibishev), Pscov, Yaroslavl - Central Russian dialects. Nizhny Tagil, Novo sibirsk, Omsk, Tomsk, Sverdlovsk, Chelyabinsk, Perm are included into the Ural-Siberian group. Leningrad, as it is known, doesn't belong to the zone of functioning of any local dialects.

Analising the speech of people living in these cities we have an oppor tunity to observe the effect of local dialects and popular speech on the standard language, as well as to find out the correlation between their fre quency and the level of education (secondary in complete higher) and also the profession (philologist - or not). The speakers were chosen from the natives of the city, from 18 to 60 years of age, which had secondary or higher 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 recor ded 21% (126 people) of the total quantity of the subjects had higher educati on, 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 analized mainly by ear. The results has showed 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 reproductions of popular features only as about a tendency: that is the lack of occlusion during the pronunciation of the affricate /c/, the lack of dissimi lation in the consonant cluster in the word /l´ixko/ read as /l´ikko/ 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 frequency of subnormal features of pronunciation and the level of education: the higher the frequency of the popular and dialectal elements is - the loiser is the level of education. It's re markable 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 dialectal, features, for example the pronunciation of the fricative [ɣ] instead of the normally occlusive [g]. In the speech of the subjects from all the towns, exept Southern, popular features are 2-3 times more frequent than dialectal ones. The simplification of the final consonant groups such as /s´t´/ - /s´/, /z´n´/ - /s´/, /pav´erxnas´/, /žɪs´/ is widely spread everywhere.

In all the cities, exept Leningrad, the pronunciation of students is to a larger extent more orphoepic 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. Heuce, 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 tipical deviations from norm are presented in Table. The percentage of philologists and subjects of other professions grouped ac cording to the regions is the following: in the North-Russian cities the philologists comprise 5% 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%.

Cities	Profession	Deviations of pronunciation [e]	[o]	/č/-/š/ /c/-/s/	/s't'/-[s'] /žn'/-[s']	[ɣ]	/k/-/x/
North-Russian	philologists	7	6	22	58	-	-
	others	16	7	26	92	-	-
Uralo-Siberian	philologists	3	3	6	17	-	-
	others	3	4	12	25	-	-
Middle-Russian	philologists	9	-	20	35	-	-
	others	8	-	13	43	-	-
South-Russian	philologists	5	-	11	32	12	45
	others	9	-	10	23	26	52

On the basis of the given data the following conclusives can be made:

1. The deviation from norm in the students' speech is the result of the in fluence of the dialect, in which region the city is situated, and also of popular speech, which is locally not limited. Thereby the frequency of the dialectal features,as a rule, is lower than popular ones,exept the case with the South-Russian cities where fricative [ɣ] 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 casuality and passivness of articu lation, are more frequent in the speech of non-philogists. In North-Russian cities the students use popular elements more frequently than in other regions.

4. Dialectal features of pronunsiation are to a larger extent peculiar to the

speech of non-philologists.

5. The percetage of the appearance of /e/ in the unstressed position is relatively small inspite its territory wide-spread character, wich proves the gradual establishment of national - wide choice of /i/ in this position. Thus we suggest that the further development of the orphoepic standard will draw nearer with the popular elements and will come to the spreading of the pronunciation of affricates, especially /č/ without the occlusive phase and to the sim plification of the final combination /s't'/ into /s'/. A gradual penetration of the super-frequent 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.

Summery. The speech of 588 people living in 20 cities of Russia is analized 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 exept 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 exept 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 Hammarberg

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 an individual'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, pharyngalization, raised larynx etc., and phonatory settings, which are described partly in perceptual terms, such as falsetto, creak, 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, normal, 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 der einzelnen Sprachen"; [7, 246-251], 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 distribution of such 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 c. 60 speakers (men and women) from Gällivare-Malmberget, Göteborg, 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 types 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 have 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 (FFDA) 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 Göteborg 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 standard 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 uvularization which is associated with the occurrence of uvular [R] or (in medial and final position) a central or back vowel as allophones of the frequent phoneme /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
[1] ELERT, C.-C. (1983), "Forskning om tal, ljud och hörsel i en humanistisk omgivning" ,*Tal ljud hörsel* (L. Nord & P. af Trampe, eds.), , 25-36, Stockholm: Institutionen för lingvistik.
[2] ESLING, J. H. (1978), *Voice quality in Edinburgh: a sociolinguistic and phonetic study*, Ph.D. dissertation, University of Edinburgh.
[3] FOLDVIK, A. K. (1981), "Voice quality in Norwegian dialects", *Nordic prosody II* (T. Fretheim, ed.), 228-232, Trondheim: Tapir.
[4] GOBL, C. (1989), "A preliminary study of acoustic voice quality correlates", *Speech Transmission Lab., QPSR*, Stockholm: K. Tekniska högskolan, 4/1989, 9-22.
[5] HAMMARBERG, B. (1986), *Perceptual and acoustic analysis of dysphonia.* Stockholm/Huddinge: Dpt. of Logopedics and Phoniatrics, Karolinska institutet.
[6] HONIKMAN, B. 1964. "Articulatory settings", *In honour of Daniel Jones*, London: Longmans, 73-84.
[7] JESPERSEN, O. 1920 [1912]. *Lehrbuch der Phonetik*, Leipzig: Teubner.
[8] JOHANSSON, I. (1978), *Funktionella aspekter på satsintonationen i svenska.* With a summary in English. (Umeå studies in the humanities 16), Umeå: Universitetsbiblioteket.
[9] LAVER, J. (1980), *The phonetic description of voice quality*, Cambridge: Cambridge U.P.
[10] PITTAM, J. (1987), "Listeners' evaluation of voice quality in Australian English speakers", *Language and speech*, 30, 99-113.

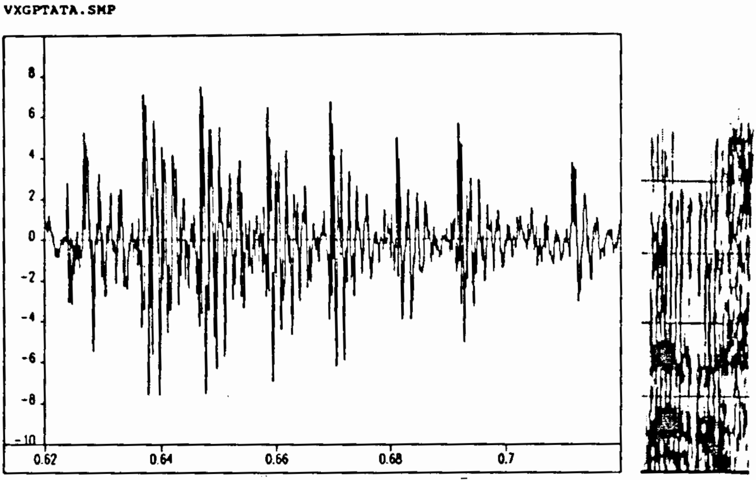


Figure 1. Waveform of [a(ʀ)] and spectrum of [aʀe] in the word *tätare* pronounced by the male Växjö speaker GP.

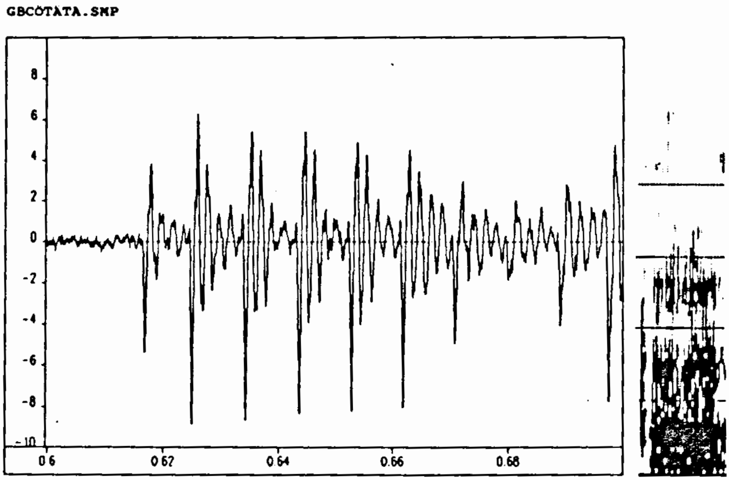


Figure 2. Waveform of [ar(e)] and spectrum of [are] in the word *tätare* pronounced by the male Göteborg speaker CÖ.

ETUDE SUR LA PERCEPTION DE L'"ACCENT" REGIONAL DU NORD ET DE L'EST DE LA FRANCE

Fernand Carton
Institut de Phonétique, 23 Bd Albert 1er, 54000 Nancy
Robert Espesser
Institut de Phonétique, URA 261, 29 av. Robert Schuman, 13 621 Aix-en-Provence
et 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 permutée 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 pénultième pour la perception d'un "accent". Un second test vise à juger de l'importance la courbe mélodique dans la perception de l'"accent". Les phrases prononcées en français standardisé et les phrases avec "accent" patoisant ont été synthétisées avec un 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 a l'intonation *traînante* 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 les expressions populaires de: "il n'a pas d'accent", "il a 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 particulière d'un seul phonème, par exemple la diphtongaison d'une voyelle ou la prononciation d'un /R/, ou les indices peuvent se répartir sur un domaine *plus vaste* de la phrase, voire sur la phrase toute entière. De plus, une combinaison d'indices *non* spécifiques peut devenir une marque d'une région particulière. Enfin, la réaction 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 voix à accent marqué peuvent être "débarrassées" 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 à "accent" marquée régionalement. Ce type d'études n'est pas seulement utile pour compléter notre connaissance sur les accents régionaux et les productions véritables, 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 démuni pour expliquer les réactions des auditeurs.

I. RESULTATS D'EXPERIENCES PRELIMINAIRES

Des études 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 clausules (trois dernières syllabes) intonatives et temporelles, indices possibles d'identification socio-géographique, le point clé du décodage se situant sur l'avant dernière syllabe du groupe. La rupture mélodique et sonie corrigée [1,2,3], la durée vocalique (durée accrue de l'avant-dernière -pénultième- syllabe calculée par rapport à la durée moyenne des syllabes accentuées [4]), et l'énergie et la durée consonantiques dans cet ordre ont montré qu'elles jouaient un rôle primordial (voir aussi [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 mélodique). La permutation de la pénultième de phrases à "accent" régional dans des phrases standardisées est-elle un indice nécessaire et/ou suffisant pour déclencher chez l'auditeur la perception d'un accent régional? (Test 1). La mise à plat des variations de Fo (suppression de toute rupture mélodique) nuit-elle réellement à la perception d'un "accent" régional? (Test 2). Notre choix s'est porté sur l'étude de phrases où l'"accent" régional est fortement marqué, afin de faciliter les tests de perception.

2. CORPUS ET LOCUTEURS

Lors d'un enregistrement préliminaire, deux locuteurs, un conteur patoisant de l'Est de la France, de la région de Nancy (Lorraine romane) (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" patoisant 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 pupitre". La pénultième de toutes les phrases comporte une syllabe ouverte commençant par la consonne sourde /p/ et suivie de la consonne sourde /p/, et toutes les voyelles du Français possibles dans cette position sont représentées. Le choix d'un contexte sourd a été choisi afin de simplifier le problème de l'extraction de la voyelle pour l'épreuve de permutation. Nous aurions voulu fixer complètement 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 naturelles satisfaisant ces conditions. Lors d'une première écoute, on a également noter 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 subséquent par d'autres locuteurs patoisants (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 soin au locuteur de les marquer régionalement, et deux fois en français standardisé. Parmi les 176 phrases (22 phrases * 2 styles * 2 répétitions * (L1 + L2)), 10 phrases (cinq 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 parisienne, et 10 auditeurs étrangers parlant français et vivant actuellement à Paris, tous étudiants à l'Institut de Phonétique de Paris III (niveau licence).

TEST I: LA PENULTIEME

Pour tester de l'importance de la pénultième, 4 types de phrases ont été présentées aux auditeurs: les 10 phrases standardisées ("Ce n'est pas mon pupitre": SS), les 10 phrases correspondantes à "accent" patoisant (*"Ce n'est pas mon pupitre"*: PP), les phrases standardisées où la pénultième a été permutée, après une 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 *pu*pitre": Sp), et les phrases régionales où la pénultième a été permutée avec la syllabe correspondante des phrases standardisées (*"Ce n'est pas mon* pupitre": Ps). Chaque phrase a été présentée deux fois dans un ordre aléatoire.

Le tableau ci-dessus indique la **durée** comparée des mêmes voyelles prononcées dans les phrases selon deux styles, l'un patoisant et l'autre standardisé. Les chiffres confirment l'allongement considérable, mais variable selon les phrases, de la pénultième des phrases avec "accent" régional.

L1	/a/	/e/	/in/	/eu/	/on/	Moy
SS	165	153	196	150	199	172
PP	271	254	386	286	371	314
%	+39%	+66%	+96%	+90%	+86%	+82%
L2	/a/	/e/	/i/	/u/	/ou/	Moy
SS	170	215	156	156	223	185
PP	289	307	263	290	274	284
%	70%	42%	68%	85%	22%	53%

Tableau 1: Durée de la pénultième (en msec) dans les phrases standardisées (SS) et les phrases patoisantes (PP) prononcées par le locuteur de Nancy (L1) et celui de Tourcoing (L2) et pourcentage d'augmentation de la durée.

Les courbes de **Fo** des phrases patoisantes indiquent des déviations importantes par rapport aux courbes mélodiques 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, suivie 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 contraste avec le schéma final descendant des phrases standardisées.

RESULTATS

Les auditeurs ont eu à juger (jugement forcé) si chaque phrase entendue possède "pas d'accent" (noté 0), "peu d'accent" (noté 1) ou "un fort accent" (noté 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 99% des phrases standardisées sont perçues comme n'étant pas ou peu marquées par les sujets français, mais seulement 85% par les sujets étrangers. Dans 15 % des cas, les étrangers perçoivent comme fortement marquées des phrases standardisées! (contre 1% des cas pour les sujets français). Par contre, les sujets étrangers perçoivent comme sans accent 11% des phrases patoisantes (contre 2% des sujets français). La plupart de ces étrangers se plaignent du manque de méthodes mises à leur disposition pour améliorer leur propre "accent" étranger, et des tests de ce genre semblent confirmer leur perception floue de la norme.

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 extraite 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 nasalisée avec l'accent régional), de la voyelle /e/ (diphtonguée dans les phrases marquées régionalement) et de la voyelle postériorisée qui est le plus efficace. Des tests en cours permettront de quantifier l'apport de chaque "écart" de prononciation par rapport à la norme et d'expliquer les cas où la permutation s'est révélée inefficace.

Tableau 2a	0	1	2
SS	72%	27%	1%
PP	2%	24%	74%
Sp	14%	23%	63%
Ps	39%	34%	27%
Total	32%	27%	41%
Tableau 2b	**0**	**1**	**2**
SS	59%	26%	15%
PP	11%	35%	54%
Sp	14%	38%	48%
Ps	28%	43%	29%
Total	28%	35%	36%

Tableau 2: Résultats du Test 1 sur les auditeurs français (2a) et étrangers (2b). Sp représente les phrases standardisées où le pénultième a été permutée avec le pénultième de la phrase à accent correspondante. 0, 1 et 2 correspondent aux 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

normale en une phrase marquée. Dans la majorité des cas (61%), la phrase est toujours perçue comme étant peu (34%) ou fortement marquée (27%) et la permutation n'est efficace que dans 39% des cas. L'efficacité du changement varie en fonction des phonèmes restants.

TEST 2

Le Fo des phrases de L1 et L2 a été mis constant et égal à la fréquence fondamentale moyenne de la phrase.

Tableau 3a	0	1	2
SS	58%	36%	6%
PP	5%	25%	70%
Total	31%	30%	38%
Tableau 3b	**0**	**1**	**2**
SS	43%	34%	25%
PP	25%	35%	40%
Total	34%	34%	32%

Tableau 3: Résultats du Test 2 sur les auditeurs français (3a) et étrangers (2b).

Les résultats ne sont pas faciles à interpréter car les phrases standardisées à Fo plat ont tendance à être percues comme marquées: seulement 58% d'entre elles sont perçues comme sans accent par les sujets français. Moins de phrases patoisantes sont perçues comme ayant un fort accent (de 74% à 70%) et plus de ces phrases (de 2% à 5%) sont perçues comme sans accent, ce qui confirme la contribution de la rupture mélodique comme indice. Les résultats des étrangers deviennent très aléatoires: 25 % des phrases standardisées sont perçues comme ayant un accent très marqué et 25 % des phrases patoisantes sont perçues comme sans accent.

CONCLUSION

Le premier test a confirmé à la fois le rôle important de la pénultième dans la perception d'une marque d'une région particulière, et l'incidence d'autres facteurs. Le second test a montré que l'absence de rupture mélodique 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" ([6] 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, aussi grand soit-il, ne pourrait permettre d'apporter une réponse définitive au problème de la combinaison des indices non spécifiques qui deviennent une marque. Chaque hypothèse tirée de l'analyse d'un corpus doit être testée par des expériences de manipulations des différents indices découverts par l'analyse. Avec l'avènement relativement récent de méthodes efficaces (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 et la technique devance notre savoir: saurons-nous en tirer pleinement profit?

REFERENCES

[1] CARTON, F., & LONCHAMP, (1979), "Expérience sur la reconnaissance des traits intonatifs dialectaux (analyse multidimensionnelle)", *8e Congrès International des Sciences Phonétiques*, Leeds, 1975 et Verbum, Nancy, 88-99.

[2] CARTON, F., (1980), "L'accentuation dans le français dialectal du Nord de la France", dans <u>L'Accent en Français Contemporain</u>, *Studia Phonetica.*

[3] CARTON, F., (1981), "Les clausules comme variations rythmiques, exemples de deux dialectes français", dans Hommages à Georges Faure, *Studia Phonetica*, Vol. 2, Didier, Montréal, 78-85.

[4] CARTON, F., (1989), "La structuration temporelle dans les français régionaux du Nord-Est", dans *Mélanges de Phonétique Générale et Expérimentale offerts à Pela Simon*, Trav. de l'Institut de Phonétique de Strasbourg.

[5] CARTON, F., (1983), "A la recherche d'intonations régionales", *Actes du 17e Congrès International de Linguistique Romane*, Aix-en-Provence, 249-257.

[6] CARTON, F., ROSSI, M., AUTESSERRE. et LEON, P., (1983), *Les accents du Français*, De Bouche à Oreille. Hachette Français Langues Etrangères.

[7] KOJIMA, K., (1988), "Formes d'intonations en Français parlé", *The study of Sounds*, Vol. 22, Tokyo, Japon.

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 knows well are shorter than the same words addressed to a hearer whom the speaker has 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 have to do with its wider linguistic and paralinguistic context, and 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 on subsequent mention [4,5,6]; words are longer when they occur in sponaneous discourse than when they are produced by the same speaker reading back a transcript of the same discourse [5]. 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 evidence is found in the literature on the processing of spontaneous speech (see, for example, [5, 8, 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 with each other's speech habits). In 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 markedly 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 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 operated more often in mothers' speech to their children than in their speech to an adult visitor. In their experiment addressee 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 the dependent variable studied by Shockey and Bond was phonological rule application rather than word duration, it is not implausible that the two variables might be subject to similar influences, and indeed a further study of motherese [12], in which addressee age and familiarity were similarly confounded, revealed that words addressed to children were shorter (as well as less intelligible) than those addressed to adults.[3]

2. METHOD

The spontaneous speech samples which were used in the current study were collected using the so-called map task [13], which involves pairs of speakers, each of whom has a map. One speaker, the Instruction Giver, has a route marked on his or her map, while the other, the Instruction Follower, has no route. The speakers are told that their goal is to reproduce the Instruction Giver's route on the Instruction Follower's map. Neither speaker can see the other's map, and in the version of the task described in this paper, the speakers were prevented from seeing each other by the presence of a screen. The maps are not identical in every respect [1] and speakers are told this explicitly at the beginning of their first session. It is, however, up to the speakers to discover exactly how the two maps differ; they are encouraged to ask as many questions as necessary in order to achieve their goal. The task has been used extensively to study speakers' discourse strategies and is considered by experimenters and subjects alike 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 members of the other pair in their quadruple. Each subject participated in four 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 the same member of the other pair in their quadruple. Each speaker thus participated in two sessions in the Familiar condition (in which they knew their task 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 these items was identified on each of the four tapes (Spontaneous / Familiar; Spontaneous / Unfamiliar; Read / Familiar; Read / Unfamiliar); the materials were sampled at 16kHz and their durations measured using the ILS signal processing package, using conventional acoustic landmarks to identify word onsets and offsets [1]. The results presented in the next section were thus based on the analysis of 640 word tokens: 8 speakers X 20 word tokens X 2 addressees (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 [5]: spontaneous tokens were longer overall than read tokens ($F[1,152] = 28.08$, $p < .0001$).

Table 1: durations of words (msec)

Spkr	Fam Spont	Fam Read	Unfam Spont	Unfam Read
1	393	319	322	323
2	451	362	414	355
3	279	281	278	269
4	383	330	364	354
5	370	363	444	452
6	467	411	480	426
7	421	365	466	391
8	338	343	390	360
Mean	388	347	395	366

Addressee 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 discover 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 forms 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 the first pair to take part in the experiment, and that their 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 maps task. It may be that their unusual attention to detail in the task led them to adopt unrepresentative linguistic behaviours.

(3) See also Bard and Anderson (this volume).

REFERENCES

[1] KLATT, D. 1975. Vowel lengthening is syntactically detrmined in connected discourse. *Journal of Phonetics* 3, 129 - 140.

[2] LIEBERMAN, P. 1963. Some effects of syntactic and grammatical context on the production and perception of speech. *Language and Speech* 6, 172 - 175.

[3] HUNNICUTT, S. 1985. Intelligibility versus redundancy -- conditions of dependency. *Language and Speech* 28, 47 - 56.

[4] UMEDA, N. 1975. Vowel durations in American English. *JASA* 58, 434 - 445.

[5] FOWLER, C.A. 1988. Differential shortening of repeated content words in various communicative contexts. *Language and Speech* 31, 307 - 320.

[6] BARD, E.G., LOWE, A. & ALTMANN, G. 1989. The effect of repetition on words in recorded dictation. *Proceedings of Eurospeech '89*.

[7] VYGOTSKY, L.S. 1965. *Thought and Language*. Cambridge, Mass.: MIT Press.

[8] McALLISTER, J.M. 1989. *Lexical Stress and Lexical Access: Effects in Read and Spontaneous Speech.* Unpublished doctoral thesis, University of Edinburgh.

[9] BARD, E.G., SHILLCOCK, R.C. & ALTMANN, G.T.M. 1988. The recognition of words after their acoustic offsets: effects of subsequent context. *Perception and Psychophysics* 44 (5), 395 - 408.

[10] MEHTA, G. & CUTLER, A. 1988. Detection of target phonemes in spontaneous and read speech. *Language and Speech* 31, 135 - 156.

[11] SHOCKEY, L. & BOND, Z.S. 1980. Phonological processes in speech addressed to children. *Phonetica* 37, 267 - 274.

[12] BARD, E.G. 1982. *The intelligibility of speech addressed to children.* Unpublished doctoral thesis, University of Edinburgh.

[13] ANDERSON, A. & BROWN, G. 1984. *Teaching Talk: strategies for production and assessment.* Cambridge: CUP.

[14] McALLISTER, J.M., SOTILLO, C., BARD, E.G. & ANDERSON, A. 1990. *Using the map task to investigate speech variability.* Department of Linguistics Occasional Paper, University of Edinburgh.

COMPUTER DATA BASE AND ORPHOEPIC STUDIES

N.V.Bogdanova

Leningrad State University, USSR

ABSTRACT

The report deals with the description of orphoepic 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 Data base of Russian improving 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 on 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 - theoretic phonetics, Russian pronunciation and practical transcription - studies of phonetic peculiarities of sponteneous speech and results of different interferation processes, both between languages (Russian speech of non-Russians) and inside one language.

For all mentioned above aims it is very important to find out existing pronunciational variants for all

totality of Russian lexics, especially for peripherial part of lexical system (borrowings, abbreviations, complex words and so on). Up to now such studies were held on the limited material, the task of recieving recomendations 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 seria of orphoepic studies in which students of the philological department take part has been started in the Leningrad state University laboratory of experimental phonetics named after L.V. Shcherba. All studies are experimental phonetic including methods of auditory, instrumental and psycholinguistic analysis. Material in all cases is maximally complete - different Russian dictionaries: of new and foreign words, abbreviations and special lexics, frequency and derivational. In all cases the auditory material recorded by dictors- philologists whose normality of speech was tested and affirmed by special test, was stydied. Words with orphoepically difficult parts were put into phrases in identical sintagmatic positions. Auditors were the students and researchers of philological department. Auditor analysis was made mostly by experienced phonetists. Instrumental studies were made with the help of micro-computer of DVK-type (segmentation of auditory material, duration measurement, auditory serias preparation). Results in all cases are concrete recomendations in pronunciation and transcription as well as relations between found orphoepic variants. Some of these results are given below.

Among the words with complex consonant combinations those which contain combinations СТЛ, ЗДН, СТСК, НТСК, НДСК (коСТЛявый, безвозмеЗДНый, туриСТСКий, комендаНТСКий, шотлаНДСКий) were studied. Complete lists of such words were selected from the "Russian Derivational Dictionary" by D.Worth, A.Kozak and J.Johnson (New-York, 1970, further - RDD), those for which existing orphoepic recomendations (R.I.Avanesov, L.A.Verbitskaya, modern orphoepic dictionaries) were not enough or didn't exist at all, were included in experimental material.

Experiments showed that pronunciation of words with СТЛ depends on the route: in the words with routes -КОСТ-, -ХВАСТ-, -СТЛ- and -ТЛ- (коСТЛявый /stl′/, хваСТЛивый /stl′/, поСТЛать /stl′/, иСТЛеть /stl′/) all consonant complex is preserved in pronunciation; in other situations dieresa is observed - the lack of explosive consonant: счаСТЛивый /sl′/, совеСТЛивый /sl′/ and so on. Basing on the route it is easy to formalize the pronunciation rules of such words.

For words with ЗДН combination among two pronunciation variants - with dieresa /zn/ and literally /zdn/ the first is clearly prevailing (from 85% to 97% realizations for different words).

Study of words with СТСК, НТСК and НДСК combinations showed three pronunciation variants: with dieresa /ssk/ and /nsk/ assimilation in the place of origin /scsk/ and /ncsk/ and without dieresa /stsk/ and /ntsk/. The prevailing of first variant is rather considerable in all cases: from 75% (in word поСТСКриптум) to 98% of all realizations. In all other variants only full pronunciation of word поСТСКриптум (23,3%) must be taken into consideration without arguement.

Words with АЙО, АЙЕ and ОЙЕ 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 /j/, with /i/ and completely without /j/. The last variant turned out to be relevant for words with АЙО: 15% before the stressed /o/ - рАЙОн, мАЙОлика; 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 Rusian 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 Rusian 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, тЕОретический and so on) along with two-component realization (auditors fixed /io/, more seldom /eo/) the realization of combinmation as one vowel must be taken into consideration. In the latter case in first prestressed position the second component of combination - /a/, more seldom /o/ is recognized as a rule: in the second and further prestresed positions - first component /e/, more seldom /i/. The realization of stressed combination ЕО also turned out to be monovocal - in words метЕОр, тЕОрия, архЕОлог, архЕОграф.

The validity of recieved 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: слёЗНый - звеЗДНый, хулигаНСКий - арестаНТСКий, нарциССКий - нациСТСКий, мАЁвка - мАЙОр.

The newest borrowings into Russian language among which 10 cases with possible violation of Russian pronunciation norm were found are especially interesting for the studied problem. All in all 602 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 (коланхоэ, кейс, аэробика and so on). 36,6 % of word from the list may have a hard consonant before orthographic Е /брейк, икеба"на/, and 22 % - unstressed /o/ /консоме", бамби"но/, 10 % - long consonants outside a morpheme connection /сателли"т,

стеллара"тор/, in 11 % of words voiced consonants are possible at the word final /блюз, паб, и"мидж/. Last group of words was examined particulary carefully; we succeded to find out that the remaining voiced consonant is influenced by its´ phonetic character: the most frequent here are [dz] /ма"гериДЖ/, [z] /кюве"З/ and [b] /паБ/. By experiments it was proved that softening of hard consonants having no pair /хуаЦя"о/, remaining of /e/ in place of orthographic Е and Э including combinations with other vowels /бИЕна"ле, коланхО"Э, спири"чУЭл and so on/, tendency to letter by letter reading of complex consonant sequences /баСТНези"т, ЮНКТА"Д and others/ and a number of other phenomena is possible. As in all previous cases every word from the list was given orthoepic recommendations.

As a result of all mentioned and similar experimental research it became possible to clear up literary and dictionary orthoepic recommedatioms. These gained results will sufficiently add the Russian phonetic fund.

SUR LA CLASSIFICATION UNIVERSELLE DES SONS DU LANGAGE ET L'APhI

M.Gordina

Université de Léningrad, Léningrad,URSS

ABSTRACT

The paper is concerned with the principles on which universal articulatory classification of speech sounds rests. Phonetic transcription based on this classification must consider the phonetic capabilities of man. Different approaches to the classification of vowels and consonants are pointed out. Egressive pulmonic consonants are described with reference to their basic articulation, the deviations from the IPA chart are mentioned.

La division de la chaîne parlée en sons discrets relève des procédés phonologiques [8,9,10]. Il s'en suit que dans les langues de différents types on trouve des sons dont le statut linguistique n'est pas le même et qui possèdent des caractéristiques phonétiques particulières [3]. Cependant la classification universelle des sons du langage est possible grâce à l'unité des mécanismes de production et de perception de la parole. Pour établir une telle classification on doit tenir compte des possibilités articulatoires de l'homme, ce qui permet de prévoir des cases 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 cent ans à l'établissement du tableau des voyelles par H. Sweet [7] et développé par D. Jones [2]. Le tableau des voyelles cardinales de D. Jones, en forme de trapèze, qui est proposé aujourd'hui, presque sans modifications, par l'Association Phonétiques Internationale [6] en reflète plus ou moins approximativement les caractéristiques articulatoires. La forme du tableau et les points de repère (choisis d'une manière conventionnelle et qui ne coïncident avec aucune voyelle réelle [2], bien qu'étant en corrélation avec certains phonèmes de plusieurs langues) permettent de trouver la place pour n'importe quel sons vocalique et, par conséquent, tenir compte des oppositions phonologiques éventuelles. C'est pourquoi 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 ne s'agit dans la suite que des consonnes pulmonaires), 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 différents systèmes de classification (Zinder [10] en compte 13 et l'APhI [6] 11 à présent), 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 secondaires, 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 différents types de consonnes sont établies sur la base des oppositions phonologiques, comme le propose P.Ladefoged [5].

Cependant ce principe n'est pas réalisé d'une manière conséquente et en plus est discutable du point de vue théorique. Les oppositions phonologiques ne sont guère universelles. On pourrait trouver dans une langue non étudiée une opposition non prévue par la classification (p.ex. deux types de médiolinguales qui diffèrent grâce à l'organe passif, cf les postlinguales). D'autre part, il y a des types consonantiques qui ne forment pas d'opposition, p.ex. [m] et [ɱ], [ɸ] et [f], et dont l'articulation de base est différente. Le tableau de consonnes présente 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'APhI. Ce tableau modifié est présenté à la page suivante. On en a exclu les consonnes implosives: leur articulation de base est la même que celle des autres occlusives. Les consonnes éjectives peuvent être traitées ou bien comme ayant deux points d'articulation (au même titre que [k͡p] par exemple, puisque [ʔ] est classé parmi les occlusives) ou bien comme glottalisées, c'est-à-dire caractérisées par une articulation secondaire. Pour mieux illustrer les principes de classement on a introduit dans le tableau les affriquées (pour des raisons techniques quelques symboles sont omis); on a indiqué la possibilité de réaliser toutes les fricatives comme sonnantes.

Les consonnes sont classées selon l'organe articulatoire actif; dans certains cas l'organe passif doit aussi être pris en considération (p.ex. pour les labiales et les postlingua-

Classification des consonnes (d'après l'articulation de base)

mode d'obstruction \ organe articulatoire actif				labiales: bilabiales	labiales: labio-dentales	linguo-labiales	prélinguales: dorsales	prélinguales: apicales	prélinguales: cacuminales	prélinguales: rétroflexes	médiolinguales	postlinguales vélaires	postlinguales uvulaires	uvulaires faucales	pharyngales: supérieures	pharyngales: inférieures	glottales
occlusives	bruits	occlusives pures		p b		p̺ b̺	t d	t̲ d̲	ṭ ḍ	ʈ ɖ	c ɟ	k g	q ɢ				ʔ —
occlusives	bruits	affriquées médianes	à fente ronde	bʷ			tˢ	d̲ᶻ		dʐ	cṡ						
occlusives	bruits	affriquées médianes	à fente plate	pᶲ			dᶞ	t̲ᶿ	ṭš ḍž		cç						
occlusives	bruits	affriquées médianes	à fente longue	—	—	—	tʃ	d̲ʒ									
occlusives	bruits	affriquées	latérales										—	—	—	—	—
occlusives	sonnantes (nasales)			m	ɱ	m̺	n	n̲	ṇ	ɳ	ɲ	ŋ	ɴ	—	—	—	—
fricatives	bruits	médianes	à fente ronde	ʍ w	f v		s z	s̲ z̲	ṣ ẓ	ʂ ʐ	ṡ ż	x ɣ	χ ʁ		χ̌ ʁ̌	ħ ʕ	h ɦ
fricatives	bruits	médianes	à fente plate	ɸ β			θ ð	θ̲ ð̲	ṣ̌ ẓ̌		ç ʝ						
fricatives	bruits	médianes	à fente longue	—	—	—	ʃ ʒ	ʃ̲ ʒ̲									
fricatives	bruits	latérales		—	—		ɬ ɮ	ɬ̲ ɮ̲	ɬ̣ ɮ̣				—	—	—	—	—
fricatives	sonnantes	médianes	à fente ronde	w	ʋ		z	z̲	ẓ	ʐ	ż	ɣ	ʁ		ʁ̌	ʕ	ɦ
fricatives	sonnantes	médianes	à fente plate	β			ð	ð̲	ẓ̌		ʝ						
fricatives	sonnantes	médianes	à fente longue	—	—	—	ʒ	ʒ̲									
fricatives	sonnantes	latérales		—	—		l	l̲	ḷ	ɭ	ʎ	ʟ		—	—	—	—
roulées bruits et sonnantes	médianes			ʙ	—	—	—	r̲	ṛ	ɽ	—	—	ʀ	—	—	—	—
roulées bruits et sonnantes	latérales			—	—	—				—	—	—	—	—	—	—	—

les). Suivant L.Ščerba et L.Zinder[10] on a indiqué l'existence d'articulations uvulaires faucales: le son se forme au moment ou la luette se détache brusquement de la paroi pharyngale; russe dno 'fond', anglais sudden etc. On ne connaispas 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 phonemes indépendants dans certaines langues du Caucase 4

Le tableau présente une classification plus détaillée de fricatives médianes, qui sont divisées en trois groupes selon la forme de constriction: consonnes a fente ronde (p.ex. [s],[w]) consonnes à fente plate (p. ex. [θ], [ə]), 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 prélinguales.

Les vibrantes sont divisées 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 présence dans le tableau est justifiée par le principe d'universalité qui oblige à tenir compte de toutes les articulations possibles.

REFERENCES

[1] Catford,J.C.(1970),"The artculatory possibilities of man", Manual of phonetics ed.B.Malmberg,Amsterdam: North-Holl.Publ.Company.
[2] Jones,D.(1948),"An outline of English phonetics", Cambridge: W.Heffer.
[3] Kasevich,V.B.(1975),"Towards a phonological theory for (mono)syllabic languages", Abstract of papers. 8th International Congress of Phonetic Sciences, Leeds.
[4] Kodzasov,S.V.(1987), "Pharyngeal features in the Daghestan languages", Proceedings of Xlth ICPhS, 2, Tallinn, 142-144.
[5] Ladefoged,P., Halle,M. (1988),"Some major features of the International Phonetic Alphabet", Language,64, 577-582.
[6] "Report on the 1989 Kiel convention" (1989), Journal of the Internation nal Phonetic Association,19, 67-80.
[7] Sweet,H.(1890), "A primer of phonetic", Oxford.
[8] Ščerba,L.V.(1912), "Rousskie glasnye v katchestvennom i kolitchestvennom otnochenii, St-Pétersbourg.
[9] Trubezkoy,N.S.(1937), "Grundzüge der Phonologie", TCLP,7, Prague.
[10] Zinder,L.R.(1979), "Obchtchaya fonetika", Moskva: Vysch.chkola.

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 consonnants at word-boundaries. The question which arises is wether 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 consonnant and that there is no evidence of rearticulation during the closure phase.

1.INTRODUCTION
La gémination est décrite par Catford [1] comme l'enchaînement de deux articulations identiques. Les géminées peuvent se trouver à l'intérieur du mot et avoir dans cette position une valeur phonologique dans plusieurs langues. On peut aussi rencontrer des consonnes identiques à la frontière de deux mots.
En Français, les géminées apparues au XIème siècle se sont simplifiées en consonnes simples [2]. Bien que la graphie de consonnes doubles ait subsisté ou ait été empruntée (mots savants, reconstructions étymologiques), les géminées n'assurent plus de fonction distinctive 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 articulatoire? 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?
Durant la tenue de la consonne double, observe-t-on une phase de relâchement (ou de rupture de l'occlusion) accompagnée ou non d'un bruit d'explosion? Y a-t-il assimilation partielle ou totale du voisement?

-au niveau articulatoire, l'articulation d'une «géminée» correspond-elle, en termes d'appuis linguo-palataux, à l'articulation d'une seule consonne ou de deux? Dans le cas d'une tenue prolongée, peut-on observer un fléchissement de la ten-sion physiologique, ou un phénomène de réarticulation manifesté par une variation significative du nombre de contacts linguo-palataux?

-la quantité consonnantique correspond-elle à un plus grand effort articulatoire manifesté par une plus grande surface de contact par rapport à une consonne simple?

-les phénomènes observés diffèrent-ils lorque les deux consonnes sont de modes différents? Y a-t-il une différence entre une géminée où les deux consonnes sont de même mode par nature ou par assimilation?
Nous présentons ici les résultats d'une étude acoustique complétée par une étude articulatoire réalisée à l'aide de la palatographie dynamique [3], [4]. Cette méthode permet en effet de distinguer précisément tout changement, même infime, dans l'articulation, avec une synchronisation parfaite du signal acoustique.

2.PROCEDURE EXPERIMENTALE
Le corpus est constitué de 36 phrases naturelles brèves de type:

S1-S2-S3-(C)V1-#-X-V2-(C)

où: V1= /a/
V2=/i/, /a/; /u/
X= /t/; /d/; /k/; /g/; /tt/; /kk/; /gg/;/td/; /dt/; /kg/; /gk/

Ce corpus est lu par 10 locuteurs, 5 fois avec un débit de parole normal et 5 fois avec un débit rapide (ordre aléatoire des

phrases). 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.MESURES
La segmentation de V1 et des tenues et explosions des consonnes est réalisée à l'aide de l'éditeur de signal SIGNAIX [6] implanté sur MASSCOMP 5500.
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 tirée de l'examen des tracés palatographiques et de l'analyse détaillée du signal acoustique font apparaître:

4.1.1 Pour les géminées voisées
-Une tenue stable
-L'existence d'un seul mouvement articulatoire
-La persitance du voisement pendant toute la consonne
-L'absence de toute trace de réarticulation.

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 phénomènes 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 aprés l'occlusion articulatoire des traces de bruit en moyennes fréquences. Ce bruit ne peut être interprété que comme l'indication d'un contact occlusif insuffisament ferme. 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 succèderait une nouvelle progression des appuis linguo-palatins 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 /tt/ et les groupes /dt/, on constate une bonne corrélation entre les évènements articulatoires et les évènements acoustiques.
Le shéma général d'organisation des gestes

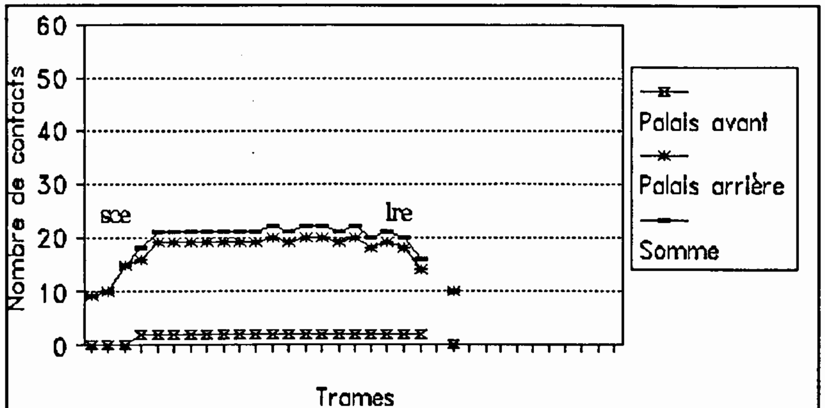


Fig.1: Nombre de contacts sur le palais avant et sur le palais arrière pour la séquence /akga/ entre l'occlusion de /k/ (sce) et le relachement de /l/ (lre).

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 apparaissent un peu plus complexes.
Nous avons pu observer à plusieurs reprises une trés légère désocclusion accompagnée de bruit. Il s'agit d'un phénomène trés 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 reprendre indiquent que non.
Il nous semble que l'on doive interpréter la trace trés 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 eu réarticulation, ce phénomène aurait dû aussi être observé sur les courbes des appuis linguo-palatins, et se manifester par exemple par un fléchissement de la tension musculaire au passage de la première à la deuxième partie de la géminée: ce qui n'a pas été le cas.

4.2.Les assimilations de voisement

Les traits de source ont souvent été considérés comme des traits redondants des traits de force d'articulation. Une consonne sourde serait forte tandis qu'une consonne sonore serait faible.
Il n'est pas possible de distinguer sur ce critère les groupes de palatales homorganiques. En effet, la mesure de l'étendue de l'appui linguo-palatal s'il constitue un des moyens d'évaluer en général la force articulatoire, est difficile à interpréter pour cette classe de consonnes en raison des débordements 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 sourdes ou assimilées sourdes sont caractérisées par un contact plus étendu de la langue au palais que les groupes de sonores correspondantes.
Par anticipation de l'articulation consonnantique, on constate généralement un phénomène d'assimilation régressive du voisement, ce qui conduit à un voisement ou un assourdissement total de la géminée. Les groupes de consonnes ainsi assimilées (sourdes ou sonores) 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 géminées.
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 géminées. L'explosion des consonnes géminées est de durée égale ou plus courte que celle des consonnes simples. (cf.Fig.2)

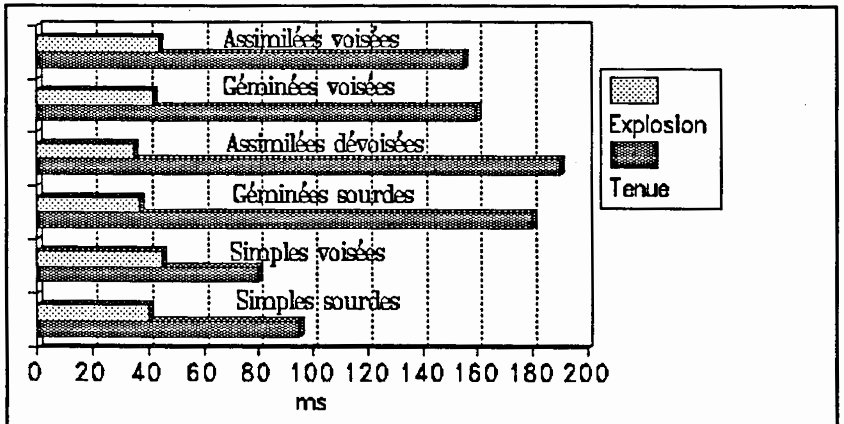


Fig.2: durées de l'occlusion et de l'explosion pour chaque mode et chaque type de gémination.

4.5.Influence du débit

L'influence du débit constatée par Wocjik [8] qui voudrait que le débit rapide entraîne une hypo-articulation n'a pas été constaté ni au niveau des consonnes simples, ni au niveau des consonnes géminées. On remarque simplement une durée de tenue occlusive plus courte en débit rapide. Cependant, les consonnes simples sont moins affectées que les géminées, indépendament du mode et du lieu d'articulation.(cf.Fig.3)
Il n'y a pas d'influence du débit sur la durée de l'explosion.(cf.fig.2)

5. CONCLUSION

Les consonnes homorganiques "géminées" se comportent comme une consonne seule, dont elles ne diffèrent que par la différence de durée de leur tenue occlusive. On ne constate pas de trace de réarticulation. La force d'articulation ne présente pas de différence significative entre les consonnes simples et les pseudo-géminées. Le débit rapide a un effet plus important sur les consonnes "géminées" 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 ACCOR.

7.REFERENCES BIBLIOGRAPHIQUES

[1] CATFORD, J.C., 1977, *«Fundamental problems in phonetics»*, University Press, Edinburgh.
[2] STRAKA, G., 1964, *«L'évolution phonétique du Latin au Français sous l'effet de l'énergie et de la faiblesse articulatoire»*, Travaux de Linguistique et de Littérature II: 17-18, Université de Strasbourg.
[3] HARDCASTLE, W.J., JONES, W., KNIGHT, C., TRUGEON, A., and CALDER, G., 1989, New developments in Electropalatography: a state of the art report, *Clinical Linguistic and Phonetics, 3*, 1:1-39.
[4] MARCHAL, A., 1988, *La Palatographie*, Ed. du CNRS, Paris.
[5] TESTON, B., et GALINDO, B., 1989, «Design and development of a work station for speech production analysis», *VERBA 90, International conference on Speech Technologies*,400-408.
[6] ESPESSER, R., et BALFOURIER,O., 1988, «Un logiciel de traitement du signal sous UNIX», *Travaux de l'Institut de Phonétique d'Aix 14*, (to appear).
[7] REPP, B., 1980, «Perception and production of two-stop consonnant sequences», *Haskins Lab., SR 63/64: 177-194.*
[8] WOCJIK,R., 1979, «The phoneme in natural phonology», in *The elements: A parasession on linguistic units and levels*, 273-284, Clinical Linguistic Society, Chicago.

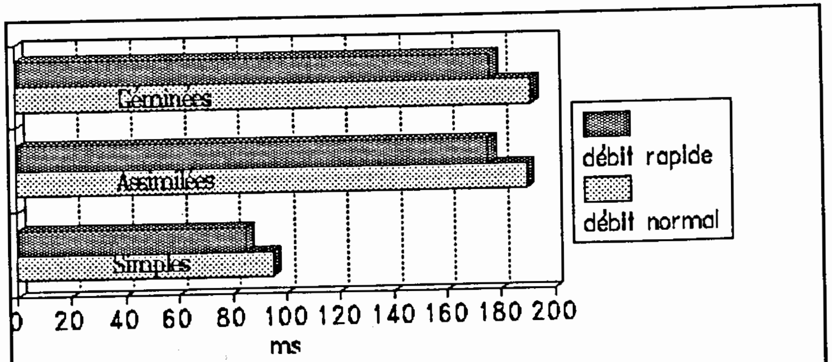


Fig.3: Influence du débit sur la durée de la tenue de l'occlusion selon les différents types de gémination.

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 absense of morphological boundaries between them. From the point of view of morphemic analysis there may be three syntagmatic types: intramorphemic, intermorphemic and mixed. Peculiarities of consonant clusters in the text in relation to the morpheme 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 a word, depending on the morphemic boundaries is very important, as such an investigation is connected with the lexico-grammatical aspect of a language.

-A Kazakh language is one of agglutinative languages and the morphological structure of words is limpid.

-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 attachment of consonant clusters to their positions in the morpheme or on the morphemic boundaries. 27 consonant clusters are intramorphemic, 15-intermorphemic, 115-mixed. Intramorphemic consonant clusters occur only in the root and are of low frequency. Their frequency is 86. Frequency of intermorphemic 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 12222.

-Medial three member consonant clusters are presented in three morphological positions: 1)on 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 intermorphemic. They are more characteristic on the boundary of two consonants of the root and the suffix and on the boundary of three suffixes 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 intramorphemic 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 9 morphological positions: 1)in the root, 2)on the boundary of the root and the suffix, 3)on the boundary of suffixes, 4)in the suffix, 5)on the boundary of the suffix and the ending, 6)on the boundary of the root and the ending, 7)on the boundary of endings, 8)in the ending, 9)on the boundary of the ending and the suffix. 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 intramorphemic, 37-intermorphemic, and 76-mixed. Their frequency is 135, 860 and 4658 respectively. Intramorphemic clusters occur only in the root of words. The most frequent intermorphemic consonant clusters are observed on the boundary of the root and the suffix, 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 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 the suffix and the ending, on the boundary of endings, and in the ending 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 5 morphological positions 1)on 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 ending, 4)on the boundary of two consonants of the root and the suffix, 5)in the root of the word. Consonant clusters on the boundary of two consonants of the root and the suffix are of high frequency. Less frequent are consonants clusters on the boundary of two consonants of

the root and the ending.
In the rest three morphological positions frequency is low. Intermorphemic 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 91%, 8%, 1% respectively.
-Final two member consonant clusters occur in two morphological positions: 1)in the root, 2)on the boundary of suffixes. They are frequent in the root of the word and they have low frequency on the boundary of suffixes. Their frequency is 81 and 2 respectively. These consonant clusters may be intramorphemic and mixed. Their percentage is 81% 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 final two member consonant clusters may be intramorphemic and mixed.
6.Both according to the dictionary and the text the most characteristic are medial two member consonant clusters of the mixed type, less characteristic are intermorphemic and the least characteristic are intramorphemic clusters.
7.Both according to the dictionary and the text medial two member consonant clusters are preferable on the boundary of the morphemes, than in the morphemes.
8.According to the dictionary intermorphemic medial two member consonant clusters are more preferable on the boundary of suffixes, than on the boundary of the root and the suffix, while according to the text they are more preferable on the boundary of the root and the suffix than on the boundary of suffixes.
9.Medial two member consonant clusters in the root of the word are more characteristic than in the suffix both according to the dictionary and the text.
10.Mixed medial two member consonant clusters according to the dictionary are more probable on the boundary of suffixes, than on the boundary of the root and the suffix, but on the boundary of the root and the suffix they are more probable than in the root of the word, and in the root of the word the consonant clusters are more probable than in the suffix. According to the text these clusters are more probable on the boundary of the root and the suffix, 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.
11.Medial three member consonant clusters are productive on the boundary of two consonants 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 three suffixes is characterised by high frequency , while according to the text that position is characterised by low frequency.
12.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 Ehlich'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 realisations 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.

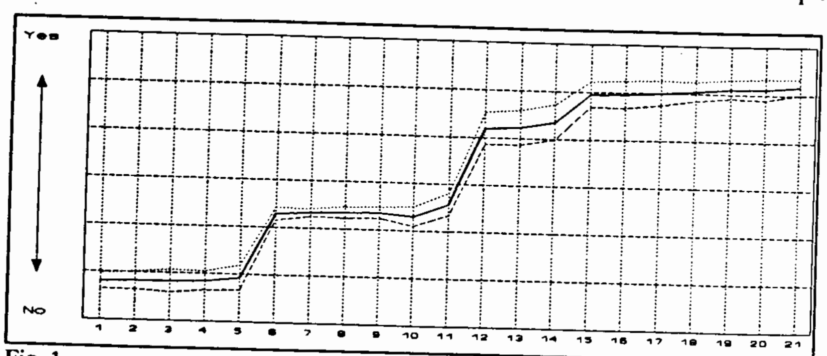


Fig. 1

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 21*6=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.

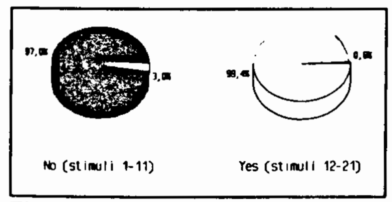


Fig. 2

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 subdivions suggested by figure 1 can be verified with different methods; cluster analysis supports the existence of four groups as well as figure 3 does.

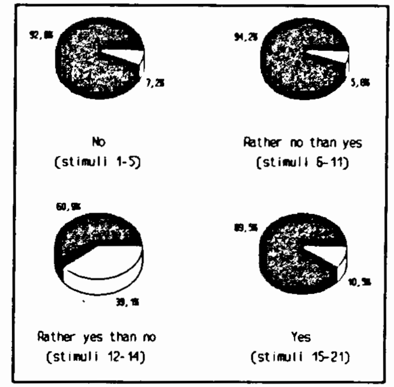


Fig. 3

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 spectre 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

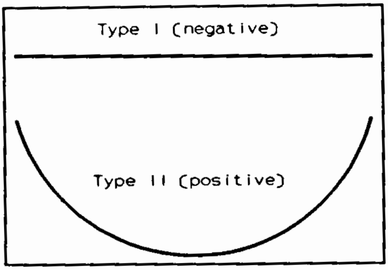


Fig. 4

variety of languages (e.g., s. [3] for Finnish). The same holds for the opposition of glottal stop and /h/ (e.g., s. [2] for English).
Figure 5 gives a general outline of the correlations between phonetic characteristics and linguistic function. It seems that in bisyllabic *hm* the stop vs. /h/ criterion takes precedence over the F0 criterion, but research on this issue is still under way.

5. CONCLUSION

hm utterances in German can, at 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), "*Interjektionen*" Tübingen: Niemeyer
[2] LUTHY, M. (1983), "Nonnative 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

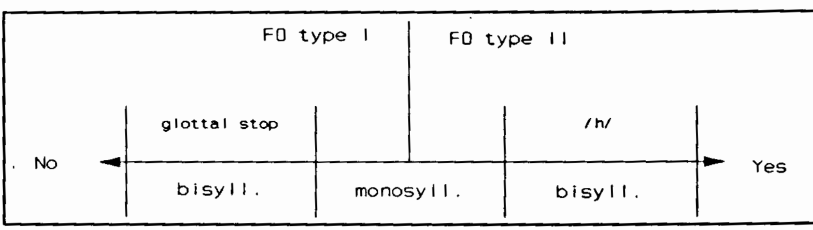


Fig. 5

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, l,w/, 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 experiences 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 being exposed to an English speaking environment has also been studied [1]. This study revealed 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 [1]), 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/ phoneme is considered in addition to /r/ and /l/, because Japanese listeners often identify some of the /r/ and /l/ sounds as /w/ [13].

2. STIMULI

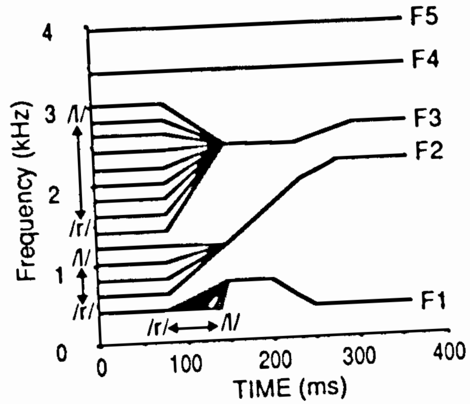


Figure 1 Schematic representation of frequency trajectories of F1 to F5 for the synthesized stimuli.

Synthesized /rait-lait/ series generated by Klatt's cascade formant synthesizer, and naturally spoken stimuli were used. Figure 1 provides a synthetic spectrographic representation of the initial CV portion, /rai-lai/, for the synthesized stimuli. The acoustic parameters for idealized "right" and "light" were derived from the naturally spoken /rait/ and /light/ uttered 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 37 combinations in total, excluding some contradictory combinations in which the F2 frequency was equal to or higher than the F3 frequency. F1 transition duration was varied from 70 ms to 16 ms in 6ms 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-lai/ part was fixed at 360 ms. The stimuli were synthesized and reproduced through 16-bit digital-analog conversion at a sampling frequency of 20 kHz and low-pass filtering with a cutoff frequency of 10 kHz. Several experiment 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 eleven blocks of ten trials and one block of one trial to make 111 trials in total. The 111 trials resulted from three randomly ordered repetitions of each of 37 stimuli. The inter-trial interval was 2 seconds, and the inter-block interval was 8 seconds. The block start signal was a beep sound recorded 2 seconds prior to the beginning of each block.

Naturally spoken stimuli contained sixteen combinations of English words. Each combination consisted of three words which were different from each other only in the initial consonant, i.e., /r/, /l/, or /w/ (e.g. "red", "led", and "wed"). The forty-eight words were spoken by two native AE speakers (one female and one male) to produce a total of ninety-six stimuli. They were recorded and converted from analog to digital at a 20-kHz sampling frequency with 16-bit accuracy. These stimuli were reproduced and were recorded on digital audio tape. In each session, each of 96 stimuli occurred once in random order to make 96 trials in total, and these 96 trials were arranged in nine blocks 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 as Group JE subjects was to fulfill all the following conditions: (1) native speaker of Japanese,

(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, pre-school 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 coincidence 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 is 19 on average, and ranged from 15 to 23, that in Group JE is 20 on average, and ranged from 13 to 40, and that in Group A is 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 also 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 stimuli whose F3 onset frequencies were higher than 2000Hz as /l/ and those which were lower as /r/ (Yamada & Tohkura, 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 ($0 \leq Cs \leq 1$). The *Cn* is the averaged correct response rates across all the naturally spoken stimuli ($0 \leq Cn \leq 1$).

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 Groups 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

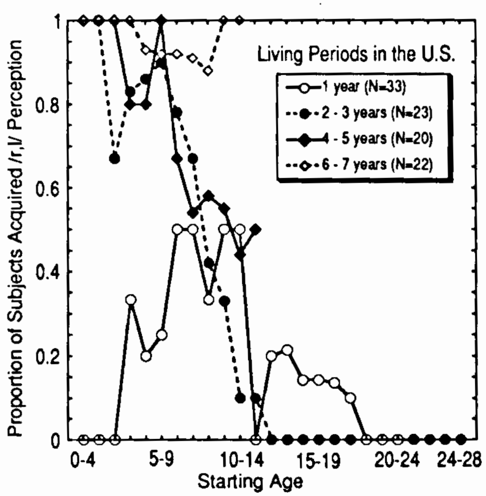


Figure 2 Capability of acquiring /r,l/ phonemes for Japanese speakers who have once been exposed to the AE speaking environment. The proportion of the subjects' number who have acquired /r,l/ perception for four groups of living periods (1 year, 2 - 3 years, 4 - 5 years, and 6 - 7 years) are represented as a function of starting age. As moving averages for each 5-year period are represented, the abscissa shows the average period.

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; $0.75 \leq Cs$, and $0.90 \leq Cn$), 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 was represented in Figure 2: subjects lived in the U.S. for 1 year, 2 - 3 years, 4 - 5 years, and 6 - 7 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,l/ 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,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

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

[2] Flege, J. (1990). "Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language," *J. Acoust. Soc. Am.* 89, 395-411.

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

[4] Mack, M. (1989). "Consonant and vowel perception and production: Early English-French bilinguals and English monolinguals," *Percept. Psychophys.* 46, 187-200.

[5] McKain, K.S., Best, C.T. & Strange, W. (1981). "Categorical perception of English /r/ and /l/ by Japanese bilinguals," *Appl. Psycholinguist.*, 2, 369-390.

[6] Miyawaki K., Strange, W., Verbrugge, R., Liberman, A.M., Fujimura, O. & Jenkins, J. (1975)"An effect of linguistic experience: the discrimination of [r] and [l] by native speakers of Japanese and English," *Percept. Psychophy.*, 18, 331-340.

[7] Mochizuki, M. (1981). "The identification of /r/ and /l/ in natural and synthesized speech," *J. Phonet.* 9, 283-303.

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

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

[10] Tahta, S., Wood, M., & Loewenthal, K. (1981). "Foreign accents: factors relating to transfer of accent from the first language to a second language," *Lang. Speech* 24, 265-272.

[11] Tahta, S., Wood, M., & Loewenthal, K. (1981). "Age changes in the ability to replicate foreign pronunciation and intonation," *Lang. Speech*, 24, 363-372.

[12] Yamada, R.A., & Tohkura, Y. (1990). "Perception and production of syllable-initial English /r/ and /l/ by native speakers of Japanese," *Proceedings of ICSLP'90*, 757-760.

[13] Yamada, R.A., & Tohkura, Y. (1991). "The effect of experiment variables on the perception of American English /r,l/ for Japanese listeners: Response categories, classification of subjects, and stimulus range," submitted to *Percept. Psychophys.*

THE REDUPLICATIVE BABBLES OF FRENCH- AND ENGLISH-LEARNING INFANTS: EVIDENCE FOR LANGUAGE-SPECIFIC RHYTHMIC INFLUENCES

Andrea G. Levitt† and Qi Wang‡

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 babbles. 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 speech has 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 thus seems to "drift" [4] in the direction of the 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 effects during the 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., [5,10]) 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 babbles of French- and English-learning infants. In the present investigation we extended our study to the rhythmic properties of those reduplicative babbles, 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 "trailer timed," according to Wenk and Wioland [14]. 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, Konopczynski [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 English-learning infants ranging in age from 8 to 12 months, found evidence for such lengthening in the babbling of some but not all of their American infants. To see 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 variable 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 infants' productions, then we would expect the French infants to exhibit more regularly-timed nonfinal syllables.

Finally, in keeping with the possibility for stress-delimited 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 [3] 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 northeastern 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 recorders using high quality microphones. Home recording sessions lasted between 10 and 20 minutes. Parents were instructed to choose a time when their child 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 taping 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, growls, emotive sounds, and vegetative noises) were digitized at 10 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 transcribed and digitized utterances, we selected all the reduplicative babbles according to our transcriptions. 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 (e.g,. fricatives 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 breathy release of certain syllables. Although nonfinal syllables could be considered to extend to the onset of the following syllable, such an alternative measure was not available for final syllables, making comparisons between nonfinal and final syllable lengths problematic. Thus, in order to avoid such difficulties, breathy releases and intersyllabic silences 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 [$t(8)=2.37$, $p=.0227$, 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 [$F(1,8)=7.48$, $p=.0256$], but no effect of early vs. late utterances and no interaction of language background and early vs. late utterances.

We measured isosyllabicity, 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 divided 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 and the late time periods. The results of an ANOVA with repeated measures showed no main effect of group (language background) and no main effect of early vs. late utterances. However, there was a significant interaction of language background and early vs. late utterances [$F(1,8)=8.402$, $p=.0199$]. As can be seen from Figure 1, the standard deviations of the utterances produced by the French infants decreased in the later stage whereas those of the American infants increased, indicating that whereas the French infants were developing more regularly timed utterances, the American infants were developing more irregularly timed productions.

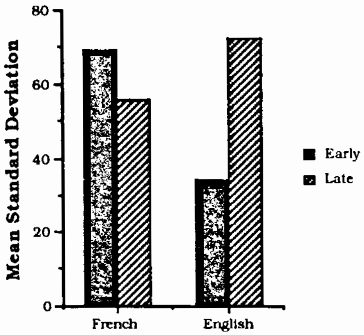


Figure 1

The mean standard deviations for nonfinal syllables produced by the French and American infants during the early and late stages of reduplicative babbling.

The percentage of "long" (four or more syllables) reduplicative babbles was calculated for each of the French and American infants. The French infants produced more long utterances (44%) than the American infants (17%). This difference was significant [$t(8)=2.901$, $p<.01$, one-tailed).

In order to see if the pattern varied over the babbling period, we recalculated the percentage of long utterances in the early 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 prelinguistic 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 regularly timed nonfinal syllables, although only in the later stage of their reduplicative babbles.

However, whereas our study of the F0 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 effects 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 [2,3] suggest 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

[1] ALLEN, G. D. (1983) Some suprasegmental contours in French two-year-old children's speech. *Phonetica, 40*, 269-292.

[2] BOYSSON-BARDIES, B. DE, HALLE, P., SAGART, L., & DURAND, C. (1984) Discernible differences in the babbling of infants according to target language. *Journal of Child Language, 11*, 1-15.

[3] BOYSSON-BARDIES, B. DE (1989) Material evidence of infant selection from the target language: a cross-linguistic study. Paper presented at the Conference on Phonological Development, Stanford U.

[4] BROWN, R. (1958) *Words and Things*. New York: Free Press.

[5] CRYSTAL, D. (1979) Prosodic Development. In P. Fletcher and M. Garman (eds.), *Language Acquisition*. Cambridge: Cambridge University Press.

[6] JAKOBSON, R. (1941) *Child Language, Aphasia, and Phonological Universals*. The Hague: Mouton.

[7] KENT, R. D. & BAUER, H. R. (1985) Vocalizations of one-year-olds. *Journal of Child Language, 12*, 491-526.

[8] KONOPCZYNSKI, G. (1986) Vers un modèle développemental du rythme français: Problèmes d'isochronie réconsidérés à la lumière des données de l'acquisition du langage. *Bulletin de l'Institut de Phonétique de Grenoble, 15*, 157-190.

[9] LOCKE, J. (1983) *Phonological Acquisition and Change*. New York: Academic Press.

[10] LEWIS, M. M. (1936) *Infant Speech: A Study of the Beginnings of Language*. New York: Harcourt Brace.

[11] MACKEN, M. & BARTON, D. (1980) The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. *Journal of Child Language, 7*, 41-74.

[12] OLLER, D. & SMITH, L. (1977) Effect of final-syllable position of vowel duration in infant babbling. *Journal of the Acoustical Society of America, 62*, 994-997.

[13] PIKE, K. (1945) *Intonation of American English*. Ann Arbor, MI: University of Michigan Press.

[14] WENK, B. J. & WIOLAND, F. (1982) Is French really syllable-timed? *Journal of Phonetics, 10*, 193-216.

[15] WHALEN, D., LEVITT, A. & WANG, Q. (in press) Intonational differences between the reduplicative babbling of French- and English-learning infants. *Journal of Child Language*.

[16] WHALEN, D., WILEY, E. RUBIN, P., & COOPER, F. (1990) The Haskins Laboratories' pulse code modulation (PCM) system. *Behavior Research Methods, Instruments, & Computers, 22*, 550-559.

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 helpfully redundant 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 show loss of 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 simultaneously learn syntax and vocabulary by listening to running speech. Word tokens in spontaneous speech are often so different from their citation forms that they have about 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, categorizing non-canonical 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 habitually 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 contexts 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 sentence, 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 predictability 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 in the second of two nearly 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 a parent spoke to his or her child and to an experimenter. Both parents of one boy and

Table 1. Distribution of Common Nouns from Speech of Twelve Parents by Addressee and Location of Referent (N = 4013; $\chi 2$ = 1371, df = 3, p < .0001)

ADDRESSEE	REFERENT LOCATION: PRESENT	PRES'T FOR SPEAKER	UNCLASS.	ABSENT	TOTAL
CHILD	1587	80	495	406	2568
(%)	(62)	(3)	(19)	(16)	
ADULT	70	7	583	785	1445
(%)	(5)	(.5)	(40)	(54)	

one girl in each of three age groups (22-24 months, 28-30 months, 34-36 months) participated. After discussing with the parent 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 engaged 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 laveliere 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 addressee 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 geographical 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-repetition, that is, in a pair of utterances 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-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 one member of a word pair. 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 5sec intervals.

Subjects and Procedure. Twenty-four native-speakers of English (6 per group) 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 homophonous identifications of the stimulus showed the expected effect of Token: first tokens were more intelligible than second tokens (57.5% v 43.5%): F_1 = 11.84, df = 1, 22, p < .005; F_2 = 3.92, df = 1, 44, p < .05; *Min* F' = 2.94, df = 1, 64, .05 < p < .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 mentioned, however, Present words are already 'Given' by extralinguistic context. The other location categories may include mentions which introduce New items. If linguistic and extra-linguistic contexts work similarly, then Present nouns should resemble co-referential or Given second mentions in having relatively low intelligibility [4, 6], whereas other categories 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-Present, 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 $F_1 = 29.14$, $df = 2, 40$, $p < .005$; $F_2 = 3.29$, $df = 2, 132$, $p < .05$; *Min F'* n. 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: *Min F'* = 4.08, $df = 1, 193$, $p < .05$; Scheffé tests 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 an 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-Present words (49%), intelligibility falls to below A-A levels. Even the Child-Unclassifiable words patterned like the unintelligible group (43%) while the Adult-Unclassifiable 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-Present words, which should be relatively unintelligible, predominate, while the clearer Child-Absent words were relatively rare, even when parents were instructed parents to produce them. Although self-repetition is not so 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

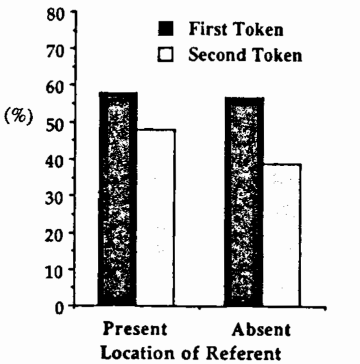


Figure 1. Intelligibility of Words in Repeated Utterances to Children (N = 96)

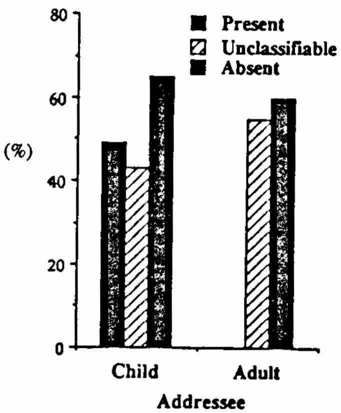


Figure 2. Intelligibility of Words by Addressee and Referent Location

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 clarity when contextual support is high, parents seem to be placing their young children at a disadvantage. To see how children might actually profit from these difficulties, it is worth considering the uses to which adults put reduced repeated word tokens. Fowler and Housum [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 that the reduced second tokens 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 lingistic context, thereby reactivating a representation of that context [3]. 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 requested via tape recordings of the words excerpted from the present corpus [1]. The children were always familiar with the nouns used 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%) overall, in this case regardless of original addressee. Moreover, originally Present words were more readily identified when the toys were visible than when they were hidden (63% correct, $N = 17$, v 36%, $N = 33$; $\beta = .279$, $t = 1.99$, $df = 48$, $p = .05$), whereas originally Absent words were less accurately identified when the toys were visible than when they were hidden (51%, $N = 30$, v 76%, $N = 14$; $\beta = -.362$, $t = -2.45$, $df = 44$, $p = .019$). 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 visible context to decode unintelligible Present words, while that context proved a distraction when they attempted to decode the 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 (SSRC HR6130) to John Laver and the first author. The authors are grateful for the help provided by the participants in the study. Reprint 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

[1] BARD, E. (1982), *The Intelligibility of Speech Addressed to Children*, Unpublished PhD Dissertation, University of Edinburgh.
[2] BARD, E., & ANDERSON, A. (1983), "The unintelligibility of speech to children", *JChLang, 10*, 265-92.
[3] BARD, E., BREW, C., & COOPER, L. (1991), "Psycholinguistic Studies on the Incremental Recognition of Speech: An Introduction to the Messy and the Sticky". Deliverable R1.3.A, DYANA, ESPRIT Basic Research Action BR3175
[4] BARD, E., LOWE, A., & ALTMANN, G. (1989), "The effects of repetition on words in recorded dictations", *Proc. EUROSPEECH '89, 2*, 573-6.
[5] FOWLER, C. (1988), "Differential shortening of repeated content words produced in various communicative contexts", *LangandSpeech, 28*, 47-56.
[6] FOWLER, C. & HOUSUM, J. (1987), "Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction", *JML, 26*, 489-504.
[7] HUNNICUTT, S. (1985), "Intelligibility vs redundancy - conditions of dependency", *LangandSpeech, 28*, 47-56.
[8] LIEBERMAN, P. (1963), "Some effects of the semantic and grammatical context on the production and perception of speech", *LangandSpeech, 6*, 172-5.
[9] NEWPORT, E., GLEITMAN, H., & GLEITMAN, L. (1977), "Mother, I'd rather do it myself: some effects and non-effects of maternal speech style", In C. Snow and C. Ferguson (eds.), *Talking to Children*, Cambridge: CUP.
[10] POLLACK, I. & PICKETT, J. (1963), "The intelligibility of excerpts from conversation", *LangandSpeech, 6*, 165-171.
[11] SNOW, C. (1972), "Mothers' speech to children learning language", *ChDev, 43*, 549-65.

ON THE PHONETIC SYSTEM EVOLUTION IN SOME ARCHAIC RUSSIAN DIALECTS

S.Knyazev and S.Pojaritskaya

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 articulatory and acoustical qualities of particular sounds (i.e the type of vowels labialization, their localization in more front or more back part of the oral cavity, and the mouth opening; the location and the type of consonantal articulation,etc.).This properties determine 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 active integrative processes connected with the deep sociological transformations during the last decades are also determined by the differences in the BA's.

We have chosen for our investigation a group of Archangelsk dialects from the Verchnyaia Toima region since they were examined more then sixty years ago by P.S.Kuznetsov [4]. We worked there in 1987 and 1990. A brief survey of these dialeçts' phonetic system is presented below.

2. VOWELS

The vowel system of the dialect is based on three-level triangle of 5 phonemes <и,э,а,о,у> which is found in a stressed syllable before a "hard" consonant and a tendency may be observed for this triangle to be used in the same consonantal context of unstressed syllables (*п'ил - п'ила́, р'ек - р'ека́, р'ат - р'аду́, н'ос - н'осу́, л'уп - л'убо́j*). However, the system is reduced in some other contexts. Thus, in the position between "soft" consonants the system of phonemes (R.I.Avanesov's "weak phonemes") if not distorted by any lexical or morphological parallels may be derived from two-level triangle of 3 phonemes /и,э,у/, in which /э/ is a result of all the non-high vowels neutralization (*п'ок-п'еч', р'ек-р'е́к'и, пр'ал-пр'е́л'и*). The same set is found in unstressed syllables (*п'ек'и́, р'ек'и́, пр'ед'и́*). The significance of stress factor in such a system (this kind of vocalic structure which is typical for the Northern Russian Dialects (NRD) is called "okanyie") is weakened as compared with the Southern "akanyie", while the factor of consonantal context (i.e. the influence of the "soft" consonant on the vowel of the preceeding syllable) becomes the decisive one.

While the preservation of the uniform sound shape of a morpheme is the most striking manifestation of "okanyie"(*в'ол, в'ол-а́, ву́-в'ол*) the influence of the following consonant (which raises the tonal features of a vowel and provokes the *o/e, a/e* alternations within a morpheme: *п'ок/п'еч', п'оку́/п'ек'и́* and *п'а́тoj/п'ет', п'ато́к/п'ет'и́*) has the opposite direction. This contradiction being inherent to the NRD phonetic system caused the distortion of the earlier relations within the system; namely the appearence of *o* and *a* in the position between "soft" consonants as a result of lexicalization and morphologization: *б'ер'о́за-б'ер'о́з'е, дома́м'и-дн'а́м'и*. A weak phonological contrast of some vowels on the basis of timbre parameters makes possible the *o-e* and *a-e* variability in a stressed position between "soft" consonants (*в'ед'о́т/в'ед'е́ц'е - в'ед'о́ц'е, стоjáл / стоjе́л'и- стоjáл'и, жэно́j / с'ем'jе́j -с'ем'jо́j*). It is especially typical for unstressed syllables where only the high vowels are opposed as the labialized and the non-labialized ones. The range of timbral variability is however restricted by the sound types which are the result of the regular phonetic changes before a "soft" consonant (*н'осу́-н'есу́, т'ану́-т'ену́*) and the system contains some intermediate slightly labialized sounds of the [e°] type as well as non-labialized ones of the [æ] type. In a position before the syllable containing [и] the vowel harmony is possible (*н'ис'и́, р'ик'и́, гл'ид'и́*) and the realizations of <о> and <а> may vary within rather wide limits ([ɵ-и] and [а-ɵ-и] respectively).

The narrow mouth opening and passive labial articulation result in a weakening of vowel distinctions and in centralization of vowels and cause the vowel variability. P.S.Kuznetsov [5] had pointed out in his analysis as a matter of fact the same manifestations of the variability in the vowel system, so we cannot reveal any essential change in this point of the phonological system under consideration.

All this show that the vowel variability should not be considered as a result of the primary system destruction under the unfluence of some other system (the Russian Literary one, for example; this variability is determined, first and foremost, by the BA properties and by the particular type of word prosodical organization in the NRD, where the word integrity is based on a consistent coordination of the sound chain units (of a vowel and a following consonant or of vowels from adjacent syllables) rather then on word stress.

3. CONSONANTS

3.1. Place and manner of articulation

The realization of the labial phonemes varies within rather wide limits. Labiodental phonemes may be realized in [в], [β], [м] (before nasals), sometimes [w] (chiefly before labialized vowels) and [ф], [π]: *на ла́πк'е, фс"о, два, врац", βо́зл'е, траβа́, ф ц"е́р'коф', годо́ф, сβоjо́* and *сπоjо́*(=своё), *wот, сла́мно, прамну́ц"ка, м нopé, лошу́шк'и*.

The voiced velar phoneme is plosive <г>: *гот, ун'аго́, огорот* (but: *бо́γу, боγа́тыj*). The ajective ending Masc.Sg.Gen. -ого may be realized as [ого] or [оо]: *другóго, н"икакóго, вос'мо́о, цно́о, ц"оо́*.

The palatal phoneme <j> may be not pronounced in the word initial position before <э>: *н'ео́м, эjо́, э́з"д"ит, эс"т", э́сл'и* and sometimes in the intervocalic position: *поэ́хала*. Epenthetic [j] may be inserted before initial <и>: *jих, jино́й*. Any consonant (but especially [т]) may assimilate the following [j]: *л"л"у* (=лью), *с"т"ихотвор'е́н"н"а, ружжа́, пла́т"т"ом, на тр'е́т"т"ом*.

Sibilants <ш>, <ж> are not

palatalized before front vowels and in the word final position: *покажы́*, *ид"о́ш*, *фшэр'с"т"е́* (=в шерсти). The only exception is the position after palatal consonants, where the palatal sibilant is found: *ра́н"ш"е*, *м'е́н"ш"е*, *бо́л"ш"е*. The long sibilants are almost always non-palatalized and may be realized as [шш], [штш], [жж], [ждж]: *эшшо́*, *шшшу́*, *поjежжа́л'и*, *уjежджа́jу*. It is worth noticing that the consonantal clusters [шт], [жд] may be represented by [штш], [ждж] as well: *жджа́л'и*, *пошчо́-то*, *од"е́жджы*. The only affricate phoneme in the dialect is realized by a number of sounds such as [ц], [ц'], [ц"], [с], [с'], [с"], [ч], [ч'], [ч"]. The lateral phoneme <л> is almost always pronounced as the "dark" [л] The exceptions are very few: *ушо́w*, *тошкл'и́*, *розм'а́w*, *до́wго*, *ку́шаw*, *спаw*, *коwо́т'ит*.
The phonemes <т>, <д>, <н>, <р> may be presented by dorsal dentals or by alveolar apicals. In the latter case a vowel after such a consonant becomes a front one: *тȫл"ко*, *тӳт*, *тӓм*, *н"е́тӳ*, *бу́дӳ*, *дру́жбу*, *βрӓц"*, *уна́с*.

3.2. Palatalisation

The so cold "soft" and "hard" consonants may be opposed in two different ways. The first one is identical with the Russian Literary Language (RLL): the consonants are contrasted on the basis of palatalisation. In this case almost every consonant may be non-palatalized or palatalized, i.e.: *м'ат'/мат*, *тк'от/кот*, *з'ат'/наза́т*, etc. The other type of the opposition is the place correlation (which is the characteristic for the eldest and the non-educated speakers). In this case the labial consonants are phonologically always "hard" but they recieve a slight palatalization before front vowels: *н"ев·э́ста*, *ф ц"е́р'коф*, *м·э́сто*, *б·э́р'ек*, *зап·ышшы́т*, *βо́сем*, *поп·э́j*. In this case the"soft" linguals are palatal: *н"ел"з"а́*, *т"е́сто*, *од"е́жджу*, *р"е́ц"ка*, etc. while the velars may be included (*лук/л"әк"*) or not included into the place opposition. In the latter case they receive a slight palatalization before front vowels and after palatal consonants: *на т"ел"о́г·э*, *в рuба́х·э*, *на бык·э́*, *д"е́фк·ы*, *в н"а́н"к·ах*, *ба́н"к·а*. It is worth mentioning that such a system is usually found in the pronunciation of the speakers whose [т,д,н]'s are apical (since the apical articulation hinders the process of palatalisation [4]).

3.3. Voiced-Voiceless Distinction

There are some indications that the "voiceless" and "voiced" consonants were earlier opposed on the basis of the {+/- tense} feature [1],[2],[3].
The voiceless plosives are aspirated: *т^{h}ут^h*, *к^{h}от^h*, *п^{h}а́р"ен"*.
The sonorants are devoiced after voiceless plosives and fricatives: *п^h·jут*, *к^{h}н̂"еj*, *п^{h}л̂оты́*, *п^{h}ра́во*, *сфоjа́*, *на п^{h}н̂"е́*, *на б·и́тфу*, *т^{h}фар"*.
In a position after a vowel or a sonorant before a vowel the voiced consonants may be pronounced instead of the voiceless and vice versa: *т^{h}о́л"ко*, *ин"т"ер"е́сно*, *н"и́жэ*, *топ·э́р"а* (=теперь), *ра́н"ш"е*.
The plosives in the word final position may have no release: *оп·э́т"*, *уро́к*, *п^{h}оп*.
The fricatives become long in the position before another consonant: *м·э̄сто*, *тош̄но*.
Voiced plosives may be realized as spirants: *иjо́т* (=идёт) or may be eliminated: *на́о*.
Sometimes a progressive assimilation takes place: *ра́н"ш"е*, *нац"н"у́*, *пал"т"о́*, *уjjо́ш* (=уйдёш), *г"г"е*(=где).
The experimental studies of the voiced/voiceless distinction [2],[3] in such dialects show that this type of correlation is entirely parallel with the tense/lax contrast in some German dialects and differs significantly from the corresponding opposition in the Finno-Ugric dialects that are neighbours of the Russian dialects in question (cf. [1]).

3.4. Kuznetsov's description

When our data presented above is compared with the description of the same dialect made in 1930 it is easy to see the points which changes most obviously: the voiceless labials <ф> and <ф'> are established; the alternation [л]/[ў] disappears; the sibilants [ш], [ж] lose the palatalization; a tendency may be observed to use more than one affricate phoneme; the palatalized labials substitute the non-palatalized ones in the word final position; the voiceless consonants substitute the voiced ones in the position before another voiceless consonant or before a pause. The {+/- tense} and the place correlations turn into the {+/- voice} and {+/- palatalization} oppositions. Nevertheless, some phonetic manifestations of the former correlations remaine: palatal articulation of the "soft" consonants, the progressive direction of the assimilative processes, the aspirated plosives, etc.

4. CONCLUSIONS

The phonetic system of the dialect evolves in a way of covergence with the RLL system, but the vowel structure remaines more stable then the consonantal one because of its less importance in a system. The most stable points of the consonantal structure are those which are determined by the BA properties and may be included into the other fonological system. The NRD system loses the most evident sound contrasts with the RLL but preserves such latent peculiarities as the apical and palatal articulations, aspiration and progressive assimilation.

REFERENCES

[1] KASATKIN, L.L; KASATKINA R.F. (1987) "The Correlation of the Tense-Lax Consonants in Some Russian Dialects and in Slavic Languages", *Proceedings XI-th ICPhS*. *Vol. 5*.
[2] КНЯЗЕВ, С.В. (1989) "Некоторые результаты экспериментально-фонетического исследования реализации противопоставления глухих/звонких согласных", *Бюллетень фонетического фонда русского языка*, *2*.
[3] КНЯЗЕВ, С.В. (1990) "Реализация противопоставления глухих/звонких согласных в некоторых германских, славянских и финно-угорских диалектах", *Congressus Septimus Internationalis Fenno - Ugristarum*, *3A*, *56-60*.
[4] КУЗНЕЦОВ, П.С. (1949) "О говорах Верхней Пинеги и Верхней Тоймы", *Материалы и исследования по русской диалектологии*, *Vol. 3*, *Москва*, *5-44*.
[5] КУЗНЕЦОВА, А.М. (1969) "Некоторые вопросы фонетической характеристики явления твердости/мягкости согласных в русских говорах", *Экспериментально-фонетическое изучение русских говоров*, *Москва*, *35-137*.

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 preaspirative 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.

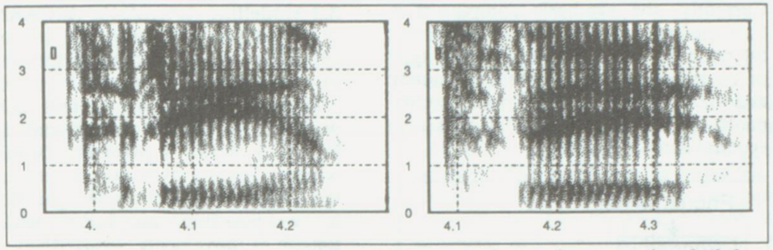


Figure 1. Spectrograms of the word [tre:] for two male speakers, speaker 1 (left) and speaker 2 (right).

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-sensitive optional 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 language of a complete recognition task. Currently, the synthesis component is formant-based. In the phoneme library, the phonemes are specified by their type of excitation and by formant frequency and bandwidth values. Certain consonants, like nasals and fricatives also have spectral zeros specified.

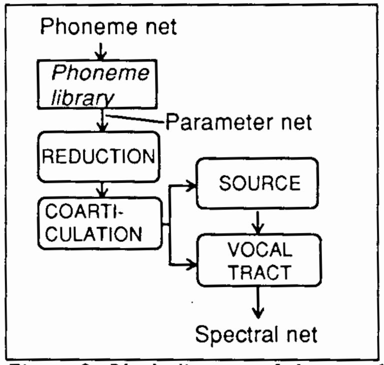


Figure 2. Block diagram of the speech production system used for reference template generation

The reduction and coarticulation components modify and expand the input phonetic network. The reduction part adjusts the targets of vowels depending on their assigned stress and their context. Since there may be more than one left or right neighbouring phoneme, it is necessary to create copies of the vowel node for all the possible contexts before applying context-sensitive formant adjustment rules.

The coarticulation component handles the transient portions at the boundaries between two phonemes. Several subphonemic states are inserted between the steady-state parts. The production parameters in these states are interpolated from the surrounding steady-state values. The number of subphonemic states in a boundary is determined by the spectral distance between the two phonemes.

The final step is to compute prototype spectra from the production parameters at each state. This is done by logarithmic addition of an excitation spectrum and transfer functions of individual formants.

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 into 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.

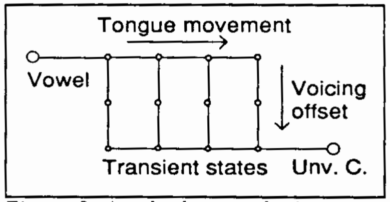


Figure 3. A sub-phoneme lattice representing varying parameter inter-timing in the transition between a vowel and an unvoiced consonant.

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 Bark-scale 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.

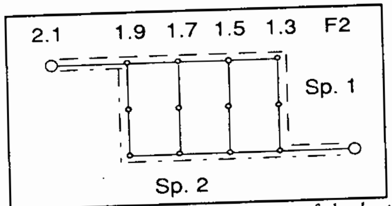


Figure 4. Results of alignment of the last part of the phrase-final [e:] The paths for speakers 1 and 2 are displayed. State values of the second formant are shown.

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. Three allowed positions of voicing offset relative to the start of the formant transitions.

Devoicing	Speaker 1	Speaker 2
Early	165	110
Late	133	160
Variable	133	111

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 recognizer, 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

[1] BLOMBERG, M. (1989) "Synthetic phoneme prototypes in a connected-word speech recognition system", *Proc. ICASSP 90*, 687-690.

[2] BLOMBERG, M. (1989) "Voice source adaptation of synthetic phoneme spectra in speech recognition", *Eurospeech 89, Vol 2*, 621-624.

[3] COHEN, M., MURVEIT, H., BERNSTEIN, J., PRICE, P., & WEINTRAUB, M. (1990), "The DECIPHER speech recognition system", *Proc. ICASSP 90*, 77-80.

[4] COHEN, P. & MERCER, R. (1975), "The phonological component of an automatic speech-recognition system", *Speech Recognition*, R. Reddy, ed., Academic Press, New York, 275-320.

[5] MARIANI, J. (1989), "Recent advances in speech processing", *Proc. ICASSP 89*, 429 - 439.

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 Electronica y Tcno. 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 technique has reached a very high performance for isolated and connected word recognition and for continuous speech recognition [1]. Applications such as voice dialing of telephone numbers 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 to be as 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. The problem is to decide *a priori* which utterance may yield a misrecognition. We propose a rejection 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.

This work has been supported by CICYT (project TIC 88-0774).

2. THE HMM-BASED RECOGNITION SYSTEM

The data were sampled at 8.091 KHz, and preemphasized with a preemphasis factor μ=0.95. Hamming windows were applied to blocks of 256 samples, with an overlapping of 64 samples. Liftered Cesptrum 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 $\pm$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 experiences, 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, HOMBRO, CODO, 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 better observe the variations of 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 some additional information, 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,

$$P_{sd} = \sum_{i=1}^{N} \log(p_i(d_i)) \qquad (1)$$

$$P_{wd} = \log(p_w(T)) \qquad (2)$$

where $p_i(d_i)$ is the duration distribution of state i, N is the number of states, T is the utterance duration, $p_w(T)$ 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 $d_i \cdot \bar{T}/T$ ($\bar{T}$ 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 that the word duration process is a gaussian process (what is basically true). P_{sd} and P_{wd} 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 a score threshold for each HMM λ 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 to 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 $\bar{x}-\alpha\sigma_x$, where $\bar{x}$ and σ_x are the log-score mean and the log-score standard deviation, obtained from the training data of a given word. The use of σ_x in the score yields a different threshold for each model λ. Moving this threshold (by the factor α) it is possible to get several rejection percentages ($RDB2$) on the testing database ($RDB2$=$RDB2(\alpha)$). 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 λ is as follows:

$$x=\frac{\log(p(O|\lambda))+\alpha_{sd}P_{sd}}{T} \quad (3)$$

In this case, the mean log-score per symbol includes the state duration log-score. State duration is included by the experimental weight α_{sd}. The optimal error rates for the different $p_i(d_i)$ distributions are: SD1) 4.58% (α_{sd}=0.7), SD2) 4.68% (α_{sd}=0.7), and SD3) 5.20 % (α_{sd}=1.7).

2) Experiment 2: the duration information is simply added to the mean symbol score

$$x=\frac{\log(p(O|\lambda))}{T}+\alpha_{sd}P_{sd} \quad (4)$$

The optimal error rates for the different $p_i(d_i)$ distributions are: SD1) 4.79% (α_{sd}=0.03), SD2) 4.79% (α_{sd}=0.03), and SD3) 5.20 % (α_{sd}=0.03).

3) Experiment 3: the same as experiment 1, but including word duration information,

$$x=\frac{\log(p(O|\lambda))+\alpha_{sd}P_{sd}+\alpha_{wd}P_{wd}}{T} \quad (5)$$

Word duration information is included as state duration in exp. 1, using an experimental weight α_{wd}. An experiment (using SD1, α_{sd}=0.7) was developed, obtaining that the error rate is an increasing function of α_{wd}.

4) Experiment 4: the same as experiment 2, but including word duration,

$$x=\frac{\log(p(O|\lambda))}{T}+\alpha_{sd}P_{sd}+\alpha_{wd}P_{wd} \quad (6)$$

The optimal error rate is 4.58% for α_{wd}=0.05 (using SD1, α_{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, α_{sd}=0.7,0.03),

$$x=\log(p(O|\lambda))+\alpha_{sd}P_{sd} \quad (7)$$

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 curve slope is higher). Also, the necessity of the temporal normalization for the threshold-based rejection is observed.

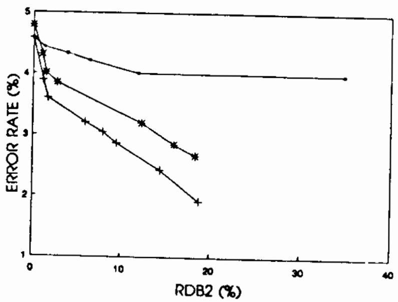


Figure 1.- Error rate vs. RDB 2 for the log-scores of experiments 1 (+) and 2 (*), and a non-normalized log-score (·).

We perform a last trial on the rejector using score (3), SD1 and α_{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 confusing words with every word of the vocabulary (3×16=48 words total). These words are divided in 3 types, according to the number and type 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* 3) as function of *RDB* 2(α). We can observe that *RDB* 3 has a good behavior for the same values of *RDB* 2 (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 uncorrect (U) words (α=3.9, *RDB* 2=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 *RDB* 2, *RDB* 3 and *RDB* 3_3 (rejection on the type 3 subset of DB3) as function of parameter α. *RDB* 2(α) has a exponential behavior, while *RDB* 3(α) has a linear one. *RDB* 3_3(α) keeps high in any case.

	R	C	U
Type-1	38.4	46.1	15.3
Type-2	55.5	33.3	11.1
Type-3	84.6	7.6	7.6

Table 1.- Percentages of R, C and U words for each type of words of DB3.

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

[1] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". *Proceedings of the IEEE*, vol. 77, no. 2, Feb. 1989.

[2] A.M. Peinado, P. Ramesh, D.B. Roe, "On the use of energy information for speech recognition using HMM". *Proceedings of EUSIPCO-90*, vol. 2, pp. 1243-1246, Sept. 1990.

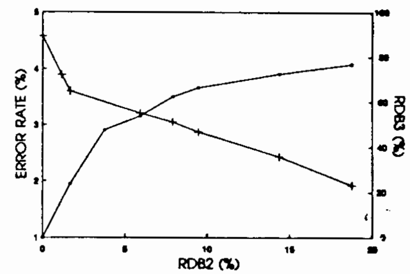


Figure 2.- Error rate (+) and RDB 3 (·) vs. RDB 2.

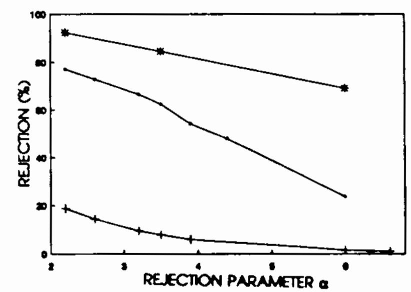


Figure 3.- RDB 2 (+), RDB 3 (·) and RDB 3_3 (*) vs. α.

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. Phonetical 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 500 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 practise, 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 achieved 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.

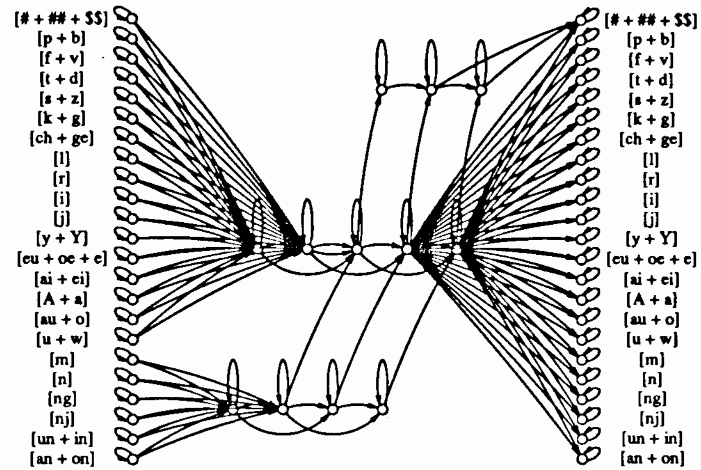


Figure 1 - Structure of the acoustical models used for the vowels, and contexts associated to the entry and exit states.

A typical model for the vowels would consist in a shared central portion representing the acoustical "target", and transitions from 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 tongue 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-rounded*, *high-mid-rounded*, *low*, *mid-front*, *mid-back*, *high-back*, *front-nasal* and *back-nasal*.

Consonantic contexts for vowel, 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 homogeneous contexts : *labial*, *labio-dental*, *dental*, *alveolar*, *palato-alveolar*, *palatal*, *velar*, /r/ and /l/. However, the nasal consonants /m/, /n/, /nj/ and /ng/ were treated as separate contexts as they may induce a nasalization of the following vowel.

Consonantic 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 consonantic contexts for consonant allophones was slightly different from that used for vowel allophones. 7 relevant contexts were defined according to acoustic features : *voiceless plosives*, *voiced plosives*, *voiceless fricatives*, *voiced fricatives*, *nasals*, /r/ and /l/.

2.2 Possible "Targets"

The inner part of the models represent the acoustical targets. Thus, in order to take into account the possible assimilation of some of the acoustic context features, several targets, representing "standard" pronunciations as well as modified ones, were modelized. The structure of the targets was carried out in order to allow the modelization of even a rather short duration of the overall sound.

Vowel targets : The following acoustical realizations were possible for the vocalic targets : *voiced, partially devoiced, partially nasalized* or *partially aspirated* (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 *devoiced* and a *partially devoiced target* (valid only for voiced consonants) and a *nasalized target.* The partially devoiced 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"* acoustic realization without any assimilation effect, a *devoiced target,* a *partially devoiced target,* and a *partially nasalized target.*

2.3 Phonological Rules

Besides the coarticulation effects between adjacent sounds treated by the allophone models, the system can handle phonological rules in order to modify the "standard" phonetic descriptions of the vocabulary words. These rules were used not to predict a 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 vocalic sound after the consonant ; 2) a voiced fricative preceded by a pause can begin with a very short schwa like vocalic sound ; 3) a succession of sounds containing a sonorant and the liquid /r/ can be realized with an epenthetic neutral schwa like vowel between them (especially in a slow speaking rate) ; 4) a voiced stop can loose 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 models (number of contexts, usefulness 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 mode). The modelization described above has then been applied (for speaker independent recognition) to several other vocabularies, recorded mainly over long distance telephone lines, by several hundreds of speakers from different parts of France, thus having different accents.

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 entry 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 decreased to 17.0 %. 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 two 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.

Structure of the models	Errors
13 contexts & 1 target	19.1 %
More contexts & targets	17.0 %
Liquids & semi-vowels shorter	16.3 %
Loops on entry & exit states	14.4 %

3.2 Efficiency of the Approach

In this section, the allophone modelization is compared to the pseudo-diphone modelization and to the word models. The standard acoustical coefficients 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 acoustical 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-diphones units on the same database.

Table 2 - Error rate obtained on several databases (for the test set) with different modelizations : Allophone models (All) ; Pseudo-Diphone units (PsD) ; and whole word models (Word).

Error rate	All	PsD	Word
Digits	**0.86 %**	1.33 %	0.69 %
Tregor	**1.00 %**	1.42 %	0.86 %
Numbers	**4.47 %**	5.68 %	——
500-Words	**8.44 %**	11.04 %	——

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 mode.

As can be seen on the above table, the results achieved by the allophone modelization are significantly better than those obtained with the pseudo-diphones units. Also, even on small vocabularies, 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 modelling 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 the same modelization to different databases led to good performances, demonstrating thus the efficiency of this new approach.

REFERENCES

[1] K. Bartkova & D. Jouvet : ***"Speaker-independent speech recognition using allophones"*** ; Proc. ICPhS 1987, Tallin, USSR, August 1987, Vol. 5, pp. 244-247.

[2] D. Jouvet, J. Monné, D. Dubois : ***"A new network-based speaker-independent connected-word recognition system"*** ; Proc. IEEE Int. Conf. ASSP, Tokyo, Japan, 1986, pp. 1109-1112.

[3] K. F. Lee, H. W. Hon, M. Y. Hwang, S. Mahajan, R. Reddy : ***"The SPHINX speech recognition system"*** ; Proc. IEEE Int. Conf. ASSP, Glasgow, Scotland, 1989, pp. 445-448.

[4] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner, J. Makhoul : ***"Context-dependent modeling for acoustic-phonetic recognition of continuous speech"*** ; Proc. IEEE Int. Conf. ASSP, Tampa, Florida, 1985, pp. 1205-1208.

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 recognition 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 $S(k)=\{S_n(k)\}$, where $n=1,\dots,N$ is the number of spectral component, and k is the number of spectral vector.

The instability of every spectral sample is:

$$E_n(k) = \sum_{g=-2}^{2} g\, S_n(k+g) \qquad (1)$$

The instability of spectral vector k is:

$$e(k)= \frac{1}{N} \sum_{n=1}^{N} | E_n(k)| \qquad (2)$$

Spectral vector $S(m)$ where $m=\underset{k}{\operatorname{argmin}}\, e(k)$ corresponds to the stable part of a phoneme and vector of instabilities $E(h)$ where $h=\underset{k}{\operatorname{argmax}}\, e(k)$ corresponds to the transitionary 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 $X=\{S(m),E(h-2),E(h-1),\dots,E(h+2)\}$ (SPCFI).

Let I, $I\geq N$ to be the number of initial features.

2.2. Dichotomization between phonemes

A linear function to discriminate between phonemes s and t is used:

$$g(s)=\sum_{i=1}^{J} W_i^{st}\,|x_i^{st}-\bar{x}_i^{s}|+Q^{st} \qquad (3)$$

where $x^{st}\in X$ represents a vector of selected features of unknown test pattern when distinguishing between s and t. Respectively $\bar{x}^s$ represents a vector of selected reference features of phoneme s, W is a vector of weights, J 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 $\bar{x}_i^s,\bar{x}_i^t$, $i=1,\dots,I$, and correlation matrixes C^s,C^t. The Gaussian distribution of features is supposed. Features are ordered according to the decrease of interphoneme distances

$$d_i^{st}=\frac{|\bar{x}_i^s-\bar{x}_i^t|}{\sigma_i^s+\sigma_i^t} \qquad (4)$$

where σ^2 is the variance. Weights W_i^{st} are calculated by using an iterative procedure to minimize the probability of misclassification p_J^{st} (J is the number of weights already defined). The procedure estimates correlations among features. Number of selected features J is computed by:

$$J=\underset{1\leq j\leq I}{\operatorname{argmin}}\, \bar{p}_j^{st} \qquad (5)$$

where $\bar{p}_j^{st}$ is expected probability of error. $\bar{p}_j^{st}$ depends on training set size, on j and on p_j^{st} and is defined from the tables in [3].

2.3. Dichotomization-based classifier

Output of an elementary dichotomie $O(s)$ is denoted by:

$$O(s)= \begin{cases} 1, & \text{when } g(s)\geq 0 \\ 0, & \text{when } g(s)<0. \end{cases} \qquad (6)$$

Respectively, output $O(t)$ is $O(t)=1-O(s)$. (7)

Here two approaches are used to get the final result:

-consistent elimination(D1): class t is excluded from the list of classes considered if $O(t)=0$, 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:

$$v=\underset{1\leq s\leq S}{\operatorname{argmax}}\, O(s) \qquad (8)$$

3. EXPERIMENTS AND RESULTS

3.1. Experimental conditions

-filter bank N=24 or N=8 (averaging 3 neighbouring) nonlinearly spaced channels;

-interval of spectral frames 10 ms;

-sample 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/,/o/,/u/,/i/ spoken by 12 males. (4800 patterns). The error rate of reference set recognition (C-examine) and "leave-one-out" recognition (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/,/u/,/i/}. 11 male speakers took part in this experiment. Reference and test sets consisted both of 220 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 4% 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 in 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/, /m'/,/n/,/n'/ in combinations nasal-vowel during training. The hard nasals were pronounced in /a/ context, the soft nasals /m'/, /n'/ - in /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/, /t/,/k/ and /b/,/d/,/g/ 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 sonant-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 males 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 descrease of recognition error rate. Dichotomization-based classifier offers the following advantages:

-inclusion of essential features only for dichotomization between phonemes;

-selection of feature set guarantying minimum probability of dichotomization error;

-immaterial influence of transmission channel because of effective application of correlations among features;

-less training set necessary to form representative references 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

[1] HARADA, T., KAWARADA,H., (1986),"High resolution frequency analysis of voices - feature extraction of /mu/ and /nu/" , *Bull. P. M. E. (I. I. T.)* - *No*. 58. *P*. 1 - 10.

[2] RABINER,L.R., SOONG,F.K. (1985), "Single-frame vowel recognition using vector-quantization with several distance measures", *AT and T Bell Lab.Techn.Journ.* -64. -*No*. 10. -*P*. 2319-2330.

[3] RAUDYS, Š., PIKELIS, V. (1975), "Tabulating of the probability of misclassification for the linear discriminant function",*Vilnius*, *Statistical methods of control*. -*No*. 11. - *P*. 81-119.

Table 1

Percentage error rates for vowels /a/,/o/,/u/,/i/

Classifier	Number of features	Examine	NS=1	NS=4	NS=11(12)
E	$I=8$	C	5.3	8.2	12.2
		L	15.6	11.5	13.2
M	$I=8$	C	0.6	0.7	1.8
		L	14.3	10.6	6.6
D1	$J\leq 4$ ($I=8$)	C	-	1.0	2.6
		L	-	3.9	2.8

NS is the number of speakers used for reference forming

Table 2

Percentage error rates for coarticulated /m/,/n/,/v/,/l/

Classifier	Method of phoneme representation	Number of features	Vowel of diphone /a/	/u/	/i/
E	SP	$I=24$	11.0	16.1	18.8
E	SPI	$I=48$	5.0	9.5	12.7
D2	SP	$J\leq 12$ ($I=24$)	2.9	8.3	10.0
D2	SPI	$J\leq 12$ ($I=48$)	1.3	3.8	6.7
D2	SPCFI	$J\leq 12$($I=144$)	-	-	1.7

Table 3

Testing of microphone change (percentage error rates)

Classifier	The former microphone	Another type microphone
E	4.0	13.9
D1	1.5	2.6

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 '*S*tatistical *Syl* labic *K*nowledge'. 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 /r/ and its voiceless allophone [ɹ̥] found after /p t k/ in words such as 'pray', 'tray', 'cray': a phoneme-based recogniser trained to recognise /r/ 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 SYLKsymbols 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 <u>within</u> 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-<u>sized</u> 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 /spr/, /str/, /skr/, /spl/, /skl/, but we do think it feasible to aim to recognise <u>the class of</u> /spr/, /str/, /skr/, 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 SYLKunits ([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 (4200 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 remained 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 recognising sub-word units. Two different units were chosen, one a phoneme-sized unit based on the segments labelled in the TIMIT corpus, and the other the SYLKunit as described above. For the former, we trained models on every phonetic category. However, in its most detailed form, the TIMIT transcription distinguishes between *the silent 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 characterised as Onset or Coda, the re-coding required decisions about syllable boundaries; as is usual, such decision were based on the *Maximal Onsets* principal according to which all intervocalic 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):

	Correct	Accurate
TIMIT LABELS.	56.6%	51.6%
SYLK-SYMBOLS.	67.9% (a) 60.8% (b)	53.5% (a) 57.7% (b)

Table 1: Recognition scores; (a) and (b) are from different HMM topologies.

It is important to compare these with results from elsewhere: the closest comparison 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 64% Correct and 53.2% Accurate. Glottal stops were ignored in their study, whereas we treat this as one of the phones to be recognised.

4.2 Comparative Evaluation: Phonetic Segments vs. SYLKunits

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 sizes and containing different amounts of information. We need to know which of the two units brings us *in principle* closest to successful word recognition. One way of doing this that we are currently investigating is first to discover which representation gives least 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 SYLKsymbols and in TIMIT phonemic symbols. Each word, in both new 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 from supplying the knowledge-based component of SYLK with both representations as partially independent sources of evidence.

5. REFERENCES

[1] ALLERHAND, M., (1987), "*Knowledge-Based Speech Pattern Recognition*", Kogan Page.
[2] BOUCHER, L.A., (1990), "Syllable-based hypothesis-refinement in SYLK", *Proc. I.O.A. 10.1*
[3] COX, S.J. (1988), "The Gillick Test - a method for comparing two speech recognisers tested on the same data", *Memorandum 4136*, RSRE Malvern.
[4] FUJIMURA, O. (1976), "Speech as concatenated demisyllables and affixes", *J.Acoust.Soc.Am.*, vol.59, p.55.
[5] GREEN, P.D., SIMONS, A.J. and ROACH, P.J. (1990), "SYLK Project foundations and overview", *Proc. I.O.A.*, vol. 10.1.
[6] HUTTENLOCHER, D.P. and ZUE, V.W. (1984), "A model of lexical access based on partial phonetic information", *Proc.ICASSP-84*.
[7] LEE, K-F, HON, W-W and REDDY, R. (1990), "An overview of the SPHINX speech recognition system", *IEEE Trans. A.S.S.P.*, vol.38.1, pp. 35-45.
[8] MILLER, D. and ISARD, S. (1984), "Aligning speech with text", *Proceedings of the Institute of Acoustics*, vol.6.4, pp.255-260.
[9] ROACH, P.J. (1990), "Phonemic transcription conventions and speech corpus design", *SYLK Working Paper No.7*.
[10] ROACH, P.J., ROACH, H.N., DEW, A.M. and ROWLANDS, P. (1990), "Phonetic analysis and the automatic segmentation and labelling of speech sounds", *J.I.P.A.*, vol.20.1, pp. 15-21.
[11] RUSSELL, M.J. et al (1990), "The ARM continuous speech recognition system", *Proc.ICASSP-90*, Vol.1, Paper S2.8, pp. 69-72.
[12] WEIGEL,W. (1990), "Continuous speech recognition with vowel-context-independent H.M.M.'s for demisyllables", *Proc.ICSLP-90*, Kobe, Japan, vol.2, pp.701-704.

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. An entire articulatory trajectory is fit in a single optimization, because an articulatory 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. The focus 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 Tukey [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 subglottal pressure and laryngeal parameters; and the work of Levinson and Schmidt [5] using a gradient search optimization to relate articulator positions to LPC 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 tube could change area through time, thus opening and closing the tube between the front and rear tubes. 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 articulatory variables was given by the model function. This function was written as an implicit relation between the formant (resonance) frequencies 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.

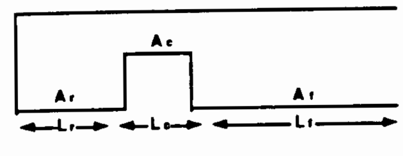


Figure 1: Modified Twin Tube

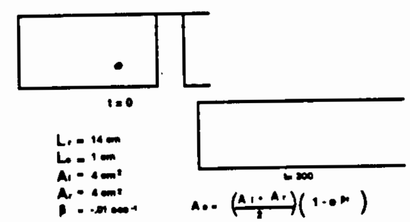


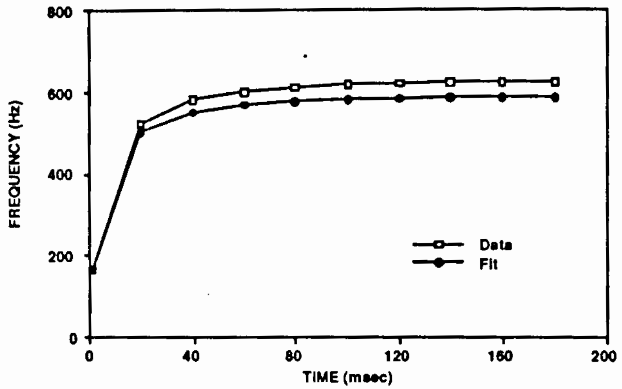
Figure 2: Articulatory Kinematic Parameters

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: rear tube length 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 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 at every iteration, many formant frequencies, corresponding to a given set of articulatory kinematic trajectories, would have had to have been 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.)



	rear area (cm2)	front area (cm2)	constriction length (cm)	rear tube length (cm)	growth factor (sec^{-1})
Data	2.	4.	2.	8.	-.01
Fit	1.684	4.178	fixed	9.7501	-.0111

Figure 3: One resonance frequency trajectory, implicit function minimization

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-01994 to Haskins Laboratories.

4. REFERENCES

[1] Atal, B. S., Chang, J. J., Mathews, & Tukey, J. W. (1978), "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer sorting technique", *J. Acoust. Soc. Am, 63*, 1535-1555.

[2] Fant, G. (1960), *"Acoustic Theory of Speech Production"*, The Hague: Mouton & Co.

[3] Flanagan, J. L., Ishizaka, K., & Shipley, (1980), "Signal models for low bit-rate coding of speech", *J. Acoust. Soc. Am., 68*, 780-791.

[4] Kuc, R., Tuteur F., & Vaisnys, (1985), "Determining vocal tract shape by applying dynamic constraints," In *Proceedings of ICASSP* (1101-1104). Tampa Florida

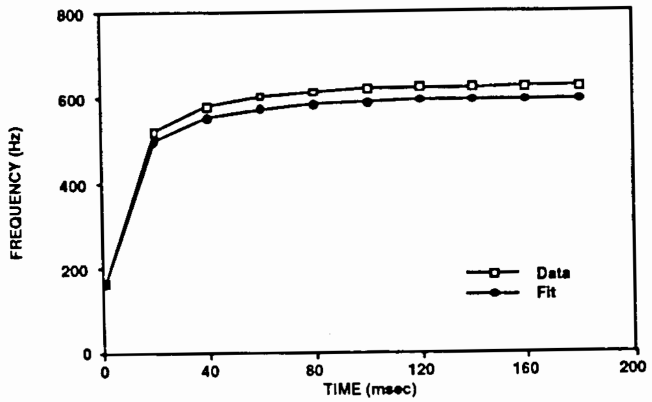
[5] Levinson, S. E. & Schmidt, C. E. (1983), "Adaptive computation of

articulatory parameters from the speech signal", *J. Acoust. Soc. Am., 74*, 1145-1154.

[6] Mermelstein, P. (1967), "Determination of the vocal-tract shape from measured formant frequencies", *J. Acoust. Soc. Am., 41*, 1238-1294.

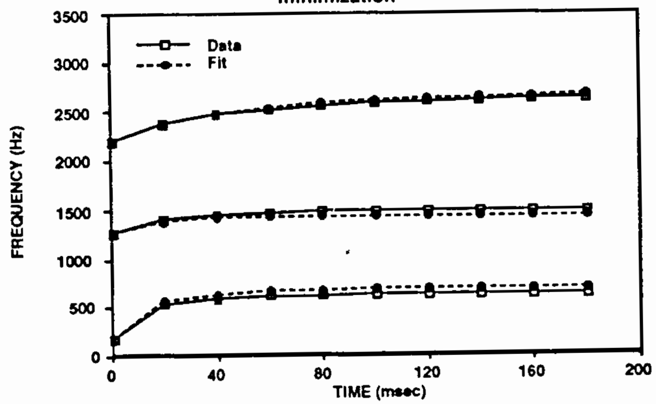
[7] Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1986), *"Numerical Recipes"*, Cambridge: Cambridge University Press.

[8] Shirai, K. & Kobayashi, T. (1986), Estimating articulatory motion from speech waves", *Speech Communication, 5*, 159-170.



	rear area (cm^2)	front area (cm^2)	constriction length (cm)	rear tube length (cm)	growth factor (sec^{-1})
Data	2.	4.	2.	8.	-.01
Fit	2.587	4.181	fixed	10.659	-.0095

Figure 4: One resonance frequency trajectory, implicit function & frequency difference minimization



	rear area (cm^2)	front area (cm^2)	constriction length (cm)	rear tube length (cm)	growth factor (sec^{-1})
Data	2.	4.	2.	8.	-.01
Fit	1.284	3.748	fixed	8.0175	-.0081

Figure 5: Three resonance frequency trajectories, implicit function & frequency difference minimization

PHONEME-LIKE MODEL OF SPEECH SIGNAL

S. Noreika and A. Rudžionis

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 metod of selecting transition and stationary frames is based on the estimation of a spectral instability function.

Let $\tilde{S}_{k,l}$ represents a set of logarithmic spectral vectors of speech signal, where k denotes the discrete time instant and $l=1,\dots,L$ denotes the number of a spectral component. The spectral instability function may be defined as follows:

$$\beta_{k,l}=\sum_{n=-n_o}^{n_o} n\,\tilde{S}_{k+n,l}\,, \qquad (1)$$

where $[-n_o,n_o]$ is the interval of spectral instability estimation, $n_o=2$.

The main segmentation function is

$$\beta_k=\frac{1}{L}\sum_{l=1}^{L}|\beta_{k,l}|. \qquad (2)$$

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 fluctuations of spectral parameters, filtering and thresholding procedures are adapted. The number of consecutive pairs "transition-stationary part" γ characterizes the extent of compression. Ideally, γ 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 PML2.

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 phonema. The clustering technique is very suitable for this purpose.

3. Comparison of test and reference patterns

One of two matrixes 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:

$$\sigma_{ij}=\begin{cases} c_{ij}, & \text{if } c_{ij}>0 \\ 0, & \text{if } c_{ij}\le 0 \end{cases}, \quad (3)$$

where $c_{ij}= d_0 - d_{ij}$ and d_0 is some constant. Several measures of coincidence between reference and test patterns were investigated, they are presented in the following section.

4. EXPERIMENTAL RESULTS

4.1. Speech material

The speech material for testing PLM1 and PLM2 was recorded by one male speaker who uttered a 100-words vocabulary 10 times. Spectral analysis of the incoming signal was carried out by a bank of 8 analog bandpass filters. All the channels were sampled every 10 ms by a 8-bits analog/digital converter. The vocabulary consisted of 794 graphic simbols, i.e. on the average one 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 like the variants V2 and V3, but the formers use 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. And finally, the seventh variant V7 is the modification of the variant V6, including the segmentation errors correction. Feature vectors for the variant V1 are represented according to PLM1, while the other variants use representation according to PLM2. The results are presented in Table 1, where N_t is the number of test patterns and ρ is the recognition error rate.

Our model ensures a high extent of compression and the number of detected phonemes γ is close to the average number of letters (7.94). Generally, the recognition error rate is inversely related to the extent of compression. The normalization of the integral similarity measure and the employment of diagonals reduce 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 per reference) and for PLM2 (memory requirement was about 50 bits per reference). The recognition results are shown in Table 2. Naturaly, VQ reduces the recognition accuracy, nevertheless, the rezults are high enough on condition that such an extent of compression is used.

4.4. PVM testing

To test this model, a 200-words 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 phonema. Some phonetic units were considered as one phonema, e.g. /p,t,k/ or /b,d,g/, so only 16 phonetic units were used. Hence it follows that memory requirement was only $4m$ 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 rezults are shown in Table 3. The model gives the promissing rezults. 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 extremal compression of speech is achieved, (2) once the database have 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 RUDŽIONIS, A. "Towards more reliable automatic recognition of the phonetic units", in this issue.

Table 1: Comparison among various variants of recognition

Variant	V1	V2	V3	V4	V5	V6	V7
N_t	9000	2700-3300				9000	
γ	9.1-7.4	7.2				7.2	
ρ,%	3.4-6.0	6.0	3.4	4.3	2.3	1.6	0.87

Table 2: PLM1 and PLM2 testing results with VQ

N_t	ρ,%		
	Without VQ	Memory requirement, bits	
		100	50
900	0.5	1.5	1.9

Table 3: PVM testing results

Number of clusters per phoneme	8	6	4
ρ,%	0.9	0.9	1.8

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 dépendent 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 /p/, /t/ et /k/ 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 valués comportant 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 ascendante 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 par le contexte. Il convient donc, dans une étape descendante du processus de décodage, de classer 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 phonémique.
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 n'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 prédiction 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.

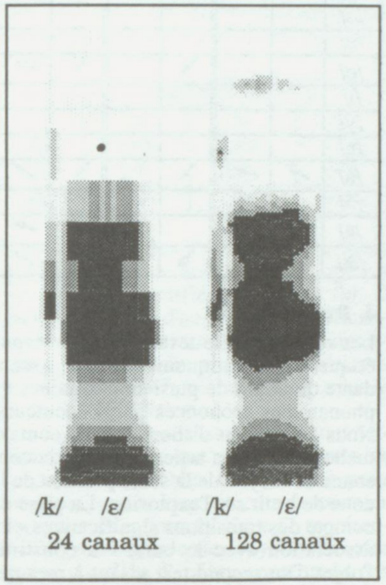


Figure 1 - La représentation spectrale au moyen de 128 canaux est nécessaire pour permettre le suivi précis des formants.

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 des valeurs des formants de chacune des voyelles pour un locuteur donné.
Pour les occlusives /p/, /t/ et /k/ des règles définissent et calculent les paramètres caractérisant l'énergie spectrale de la composante principale du 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 :

$$f(p1,v) = 2*E(p1,v) - E(p2,v) - E(p3,v)$$

qui définit la valuation de l'hypothèse correspondant à la consonne $p1$. 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.

	/p/	/t/	/k/
/a/	≤ F2-1	≥ F3+2	[F2, F2+3]
/ɔ/	[F2+3, F3-1]	≥ F3	[F2, F2+2]
/ε/	≤ F2-1	≥ F3+2	[F2, F2+3]
/œ/	≤ F2-1	≥ F3+2	[F2, F2+3]
/o/	[F2+3, F3-1]	≥ F3	[F2, F2+2]
/e/	≤ F2-1	≥ F3+1	[F2, F3]
/ø/	≤ F2-1	≥ F3+1	[F2+2, F3+1]
/i/	≤ F2+1	≥ F3+2	[F2, F3]
/y/	≤ F2-1	≥ F2+2	[F2, F2+1]
/u/	[F2+3, F3-1]	≥ F3	[F2, F2+2]
/ã/	≤ F2-1	≥ F3+1	[F2, F3]
/ɛ̃/	≤ F2-1	≥ F3+1	[F2, F3]
/ɔ̃/	≤ F2-1	≥ F3+1	[F2, F3]
/œ̃/	≤ F2-1	≥ F3+1	[F2, F3]

Les calculs de $E(p,v)$ sont effectués à partir des valeurs de la table 1. L'énergie correspond à celle du canal dont 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

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).

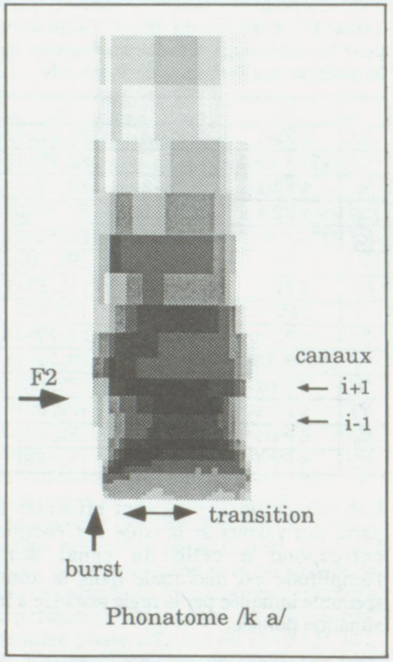


Figure 2 - Les canaux i-1 et i+1 sont utilisés pour mesurer l'évolution temporelle de l'énergie autour du formant (canal i).

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 occlusives sourdes. Seules les formes utilisées dans notre système pour l'identification du lieu articulatoire sont présentées dans cette table.

	/p/		/t/		/k/	
	F2	F3	F2	F3	F2	F3
/a/	↗	→	↘		↘	↗
/ɔ/						
/ɛ/	↗	→				
/œ/	↗	→	↘		↘	→
/o/						
/e/						
/ø/			↘		↘	→
/i/						
/y/						
/u/						
/ã/	↗	→	↘		↘	↗
/ɔ̃/						
/ɛ̃/	↗	→			↘	→
/œ̃/	↗	→	↘		↘	↗

3. RESULTATS

Les règles ont été testées sur un corpus étiqueté automatiquement (étape ascendante du DAP) de plusieurs centaines de phonatomes 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 /t/ et /k/ 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 occlusives sourdes à partir de l'explosion.

	consonne reconnue		
	/p/	/t/	/k/
/p/	**70%**	14%	16%
/t/	3%	**89%**	8%
/k/	7%	3%	**90%**

3.2. Résultats avec les transitions

Les résultats obtenus si l'on ajoute les règles caractérisant le lieu d'articulation sur les transitions de la voyelle sont donnés par la matrice de confusion de la table 4.

Table 4 - Matrice de confusion pour l'identification du lieu articulatoire des occlusives sourdes à partir de l'explosion et des transitions.

	consonne reconnue		
	/p/	/t/	/k/
/p/	**75%**	13%	12%
/t/	2%	**90%**	8%
/k/	6%	3%	**91%**

L'amélioration des résultats n'est sensible que dans le cas de la consonne /p/ qui est moins bien identifiée que /t/ et /k/. Il semble difficile d'augmenter significativement les performances de reconnaissance sans prendre en compte d'autres informations (diffusion de l'énergie sur le burst de /p/, VOT, etc.).
Les transitions utilisées (table 2) font nettement apparaître que quelques contextes sont plus favorables que d'autres pour l'évaluation des mouvements de certains formants dans notre système de représentation paramétrique (les voyelles fermées et les voyelles d'arrière constituent des environnements peu favorables). Une paramétrisation au moyen de 128 valeurs spectrales permet de mieux apprécier les transitions formantiques, mais ces informations varient parfois considérablement et sont rarement complémentaires de celles mesurées sur l'explosion [1].

4. CONCLUSION

L'identification descendante (contexte phonétique connu) des consonnes occlusives sourdes 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 multilocuteur et la discrimination d'autres phonèmes 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 pour 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

[1] BLUMSTEIN S.E., STEVENS K.N. (1979), *"Acoustic invariance in speech production : evidence from measurement of the spectral characteristics of stops consonants"*, JASA 66
[2] CALLIOPE (1989), *La parole et son traitement automatique*, Collection technique et scientifique, Masson, Paris
[3] DURAND P. (1982), *"Etude acoustique des consonnes occlusives du français commun"*, Doctorat de 3ème cycle, Université de Provence, Aix-Marseille
[4] HATON J.P. et Col. (1990), *"Décodage Acoustico-Phonétique : problèmes et éléments de solution"*, Traitement du Signal, volume 7 n°4, pp. 293-313
[5] MELONI H., GILLES P. (1991), *"Décodage Acoustico-Phonétique ascendant"*, Traitement du Signal, (à paraître)
[6] MELONI H., GILLES P., BETARI A. (1991), *"Representation of acoustic and phonetic knowledge for speaker-independent recognition of small vocabularies"*, Speech Communication, volume 10 n°2
[7] VAISSIERE J. (1987), *"Effect of phonetic context and timing on the F-pattern on the vowels in the continuous speech"*, 11ème ICPS, Tallinn, Estonia, URSS

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 vocaltract ordered by frequency. Formants carry important information about acoustic-articulatory relations, because they change their frequency and amplitude values according to different vocaltract 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 into 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 method, which allows to calculate them directly from the speech wave. They are defined by articulatory phonetics as vocaltract resonances. Formants only can be calculated indirectly via peaks or roots of the power spectrum.

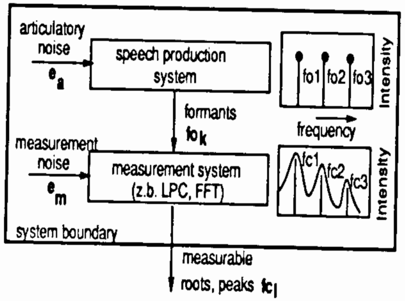


Figure 1: Formant estimation problem.

In terms of estimation theory, we can therefore formulate the following problem (also see figure 1): Suppose the peaks and roots $f c_l$ (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" $f o_k$ inside the system. So, depending on $f c_l$ we have to make an estimate for the formants $\hat{f} o_k(f c_l)$ that the estimation error $E = \hat{f} o_k(f c_l) - f o_k$ is "small".

However, this estimation process is heavily influenced by two different noise sources e_a and e_m: The errors caused by e_a have their origin in the *articulatory* system. The formant order may be confused by zeros in the vocaltract transfer function. Thus some formants are

highly damped and are not detectable. Noise source e_m causes *measurement* errors; e.g., as the fundamental frequency is superseeded 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 vocaltract resonances (=formants) is indicated by continuity.

The algorithm we present in this paper does not exclusively use smoothness criteria, because this is an oversimplification; 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.

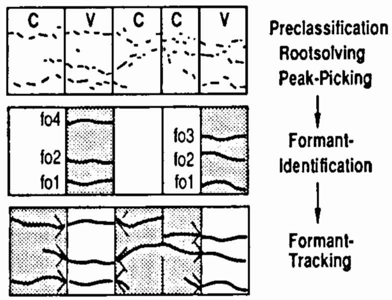


Figure 2: Schematic flow graph for formant-extraction.

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 *10m* steps. The formant candidates are determined both by peak-picking and root solving.

3.1 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_{fc} formant candidates are calculated every 10*msec*; M_{fc} is set to the number of LPC-roots minus one. Formant identification first tries to find the dominant formant regions within a segment. This is accomplished by approximating the distribution of formant candidates in V-segments by M_{fc} cluster centers with gaussian distributions. The procedure itself consists of three steps: (1) initialization of the cluster procedure, (2) calculation of cluster centers by k-means 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_{fc_l} and variance σ_{fc_l} of the formant candidate frequencies x_{fc_l} over all N_V frames i of a V-segment:

$$m_{f\,c_l} = \frac{1}{N_V}\sum_{i=1}^{N_V} x_{f\,c_l}(i)$$

$$\sigma_{f\,c_l} = \frac{1}{N_V}\sum_{i=1}^{N}(x_{f\,c_l}(i) - m_{f\,c_l})^2$$

Assuming that we already know $m_{f\,o_k}$, the speaker specific long term formant means $f\,o_k$, the centers of the cluster c_k are initialized to $m_{c_k} = (m_{f\,c_l} + m_{f\,o_k})/2$.

(2) Cluster procedure: The k-means cluster procedure is used to calculate $M_{f\,c}$ cluster centers. The resulting clusters m_{c_k}, ordered by frequency, characterize the segment specific formant frequency regions. They are defined by the mean and variance of M_{c_k} formant candidate frequencies which belong to the $k-th$ cluster:

$$m_{c_k} = \frac{1}{M_{c_k}} \sum_{x_{f\,c_l}(i)\,\in\, c_k} x_{f\,c_l}(i)$$

$$\sigma_{c_k} = \frac{1}{M_{c_k}} \sum_{x_{f\,c_l}(i)\,\in\, c_k} (x_{f\,c_l}(i) - m_{c_k})^2$$

The speaker-specific formant means $m_{f\,o_k}$ and variances $\sigma_{f\,c_l}$ are calculated by the same cluster procedure, however using a sufficient number of speech frames (about one minute).

(3) Classification: The V-segment specific formant distributions (cluster centers m_{c_k}) are used to classify the formant candidates into formants. The classification procedure maximizes the probability $p(x_{f\,o_k}(i) \mid x_{f\,c_l}(i))$ over all formant candidates $l = 1,..,M_{f\,c}$, i.e. the probability that a measured peak frequency $x_{f\,c_l}(i)$ at time i belongs to formant $f\,o_k$, when it was measured as formant candidate $f\,c_l$. The probability may be written as:

$$p(x_{f\,o_k}(i) \mid x_{f\,c_l}(i)) = \frac{1}{\sqrt{2\pi\sigma_{c_k}}}\exp(-\tfrac{1}{2}(x_{f\,c_l}(i) - m_{c_k})^2 \,/\, \sigma_{c_k}^2)$$

Applying the mean square error criterion [2] to the estimation of formant frequencies leads to the following equation: The estimated frequency $\hat{x}_{f\,o_k}$ of formant $f\,o_k$ is given by the sum of the segment specific mean frequency value m_{c_k} plus the difference of the nearest formant candidate to m_{c_k}, weighted by the maximized probability $p_{max}(x_{f\,o_k}(i) \mid x_{f\,c_l}(i))$:

$$\hat{x}_{f\,o_k}(i) = m_{c_k} + p_{max}(x_{f\,o_k}(i) \mid x_{f\,c_l}(i))\,(x_{f\,c_l}(i) - m_{c_k})$$

3.3 Formant Tracking

This part of the algorithm continues the formants of the vowel-like (V) segments into neighboured 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 corrupted formant tracks by smoothness criteria (see example in figure 2). A non-linear smoothing algorithm based on dynamic programming was choosen 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_k is measured by a cost function $c_k(l, i \mid h, i')$. It measures the deviation of formant candidates to the trajectory of formant $f\,o_k$. Assuming that the formant candidates $f\,c_l(i)$ in frame i and $f\,c_h(i')$ in frame i' belong to the trajectory of formant $f\,o_k$ at time i, the costs are given by:

$$c_k(l, i \mid h, i') = \underbrace{\frac{[\,\overbrace{|x_{f\,c_l}(i) - x_{f\,c_h}(i')|\,C_1}^{1} + \overbrace{(i - i')\,C_2}^{2}\,]}{[\,p(x_{f\,o_k}(i) \mid x_{f\,c_l}(i))\,p(x_{f\,o_k}(i) \mid x_{f\,c_h}(i')) - 1\,]}}_{3}$$

with $i' = i-1,..i-4$; $l,k = 1,..M_{f\,c}$; C_1 and C_1 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 that the formant candidates belong to formant $f\,o_k$. 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-, N_{CV} for a CV-segment respectively, has to be a minimum:

$$E = min = \sum_{k=1}^{M_{f\,o}} \sum_{i=1}^{N_{VC}, N_{CV}} c_k(l, i \mid h, i')$$

over all l, k and i'.

This equation can be elegantely 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 extending from a formant center to the left ml or the right minimum mr in the spectrum. With $s(x)$ being the log. power at frequency x, the energy to the left and right and right side $f\,e_{f\,o_k}$ of a formant center is calculated by:

$$f\,e_{f\,o_k} = \int_{x=x_{f\,o_k}}^{x=mr_{f\,o_k}\,(ml_{f\,o_k})} s(x)$$

All formant parameters are finally normalized to the speakers mean values and variances. With $f\,p_k(i)$ now being one the 21 formant parameters at time i and $m_{f\,p_k}$ and $\sigma_{f\,p_k}$ being the speaker specific means and variances of these formant parameters, the normalized formant parameters $f\,n_k(i)$ are calculated by:

$$f\,n_k(i) = \frac{f\,p_k(i) - m_{f\,p_k}}{\sigma_{f\,p_k}}$$

Expressed in filter bank terminology: The resulting parameters which are used for speech recognition are filterbank 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, u-like, o-like, a-like, ö-like, e-like und i-like)*. This task is part of an articulatory based approach for speech recognition [6].

For each articulatory category we built continuous mixture density Hidden Markov models as they are described in [4] and [6]. One version of 100 sentences was used for training, the other version was used for testing. The recognition results on *10ms frame level* are shown in Table 2. The pairs of numbers show the class specific mean recognition rates (left) and the overall frame recognition rates. The formant parameters were compared to a 16-component cepstral vector and to a 64-component feature vector as it is used in [5]. It consists of 32 mel-spectrum coefficients and differential and curvature coefficients, taking into account $\pm 40ms$ of context. 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 the 64-component vector is about 4 percent lower, but it has to be taken into account 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.

speaker	21 formant parameters	16 cepstral coefficients	64 mel differential coefficients
male1	74.7 / 84.9	66.8 / 80.3	78.4 / 86.7
male2	74.2 / 84.1	67.3 / 79.5	78.0 / 86.7
female	75.9 / 86.0	68.1 / 81.1	79.2 / 87.9

Table 2: Frame recognition rates [%] for different speakers and different feature sets.

References

[1] P. Laface. A formant tracking system towards automatic recognition of speech. *Signal Processing, North-Holland Publishing Company*, 2:113–129, 1980.

[2] F. Lewis. *Optimal Estimation*. John Wiley and Sons, New York, 1986.

[3] S. McCandless. An algorithm for automatic formant-extraction using linear prediction spectra. *ASSP*, 22:135–141, 1974.

[4] H. Ney and A. Noll. Phoneme modelling using continuous mixture densities. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 437–440, New York, April 1988.

[5] A. Paeseler and H. Ney. Phoneme-based continuous speech recognition results for different language models in the 1000-word spicos system. *Speech Communication, North-Holland Publishing Company*, 7:367–374, 1988.

[6] O. Schmidbauer. *Ein System zur Lauterkennungs auf der Basis Artikulatorischer Merkmale*. Dissertation, Fakultät für Elektro- und Informationstechnik, Technische Universität, München, 1989.

Dépôt légal - 2ème Trimestre 1991
ISBN-N°-2-85399-266-7

PROCEEDINGS OF THE XIIth INTERNATIONAL CONGRESS OF PHONETIC SCIENCES

August 19-24, 1991
AIX-EN-PROVENCE, FRANCE

ICPhS

VOLUME 4 / 5

THURSDAY AUGUST 22nd

Publication supported by :

- Centre National de la Recherche Scientifique
- Ministère de la Recherche et de la Technologie
- Délégation Générale à la Langue Française

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'Etat aux Relations Culturelles Internationales
- Ministère de l'Education Nationale, de la Jeunesse et des Sports :
  - Direction de la Recherche et des Etudes 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

SESSION 2 : Perception

SESSION 3 : Perception

SESSION 4 : **Phonétique decriptive / Descriptive phonetics**

SESSION 5 : **acquisition**

SESSION 6 : prosodie / prosody
Intonation

SESSION 7 : Prosodie / Prosody

SESSION 8 :

SESSION 9 : Phonétique appliquée : Applied phonetics

SESSION 10 : Pathologie / Pathology

SESSIONS AFFICHEES / POSTER SESSIONS

SESSION 11 : Acoustique / Acoustics

SESSION 12 : Aspects linguistiques / Linguistic aspects

SESSION 13 : Technologie / Technology

Reconnaissance automatique de la parole / Automatic speech recognition