ARTICULATORY-ACOUSTIC RELATIONSHIPS IN UNVOICED STOPS: A SIMULATION STUDY

SHINJI MAEDA

Dept. RCP
Centre National d'Etudes des Telecommunications
22301 Lannion, France

ABSTRACT

An articulatory model for consonant-vowel (CV) syllables, where C=[p, t, or k] and V=[i, a, or u], was formulated in terms of vocal-tract (VT) area function. Listener identification functions indicated that C with a high score (100 %) can be synthesized by manipulating two articulatory parameters, the "position" along the VT length and the "shape" of the occlusion. The acousic effects of these two parameters are manifested from the burst onset to the vowel transition. The consonant identity can be predicted reasonably well on the basis of the presence or absence of two spectral attributes for the burst, in a context-independent manner. Why the burst alone can predict the consonantal place? The reason is that the effects of, particularily, the shape can be manifested concomitantly on the attribute of the burst and on formant (F-) transitions of the vowel, both signaling a specific consonant. It is suggested then that the listner's processing indeed exploits cues distributed on the sound stream from the burst to vowel transitions, in which the context-independent attributes for burst may serve as an "anchor" in the identification.

INTRODUCTION

In a previous paper [1], we have described the mechanism of the VT excitation during the unvoiced stop release. The source sound, of cause, undergoes a particular spectral modification which is interpreted as a specific consonant by listeners. In the past decades, a great deal of research was accumulated in search of acoustic cues that specify place of articulation for stops. Two types of methods were employed; acoustic analysis of natural tokens to find out the acoutic correlates and examination of listener's responses to synthetic stimuli in which acoustic characteristics are systematically manipulated.

This paper describes yet another approach which consists of, first, the formulation of an articulatory model for the CV-tokens. Second, informal and formal identification tests followed to determine essential model parameters for producing the CV-syllables with a high quality. Finally, the acoustic manifestations of such parameters are examined closely by means of spectral analysis or of calculation of the VT transfer functions. If a particular manifestation can explain the listener identification of consonants, it can be considered as a good candidate for the cue actually operating in the listener's processing.

AN ARTICULATORY MODEL FOR THE CV-SYLLABLES

We assume a heavily anticipated articulation of the vowel during the preceding consonant. The VT area function defined by a piecewise-constant function is fixed to its configuration for the vowel, except in the vicinity of the occlusion, where the cross-sectional areas expand with time after release. We consider the following three different types of closure "shapes": i) Labial (L)-type; the area expansion is limited to a single section corresponding to the supraglottal closure as shown in Fig. 1a. We assume that this shape represents the release gesture by the participation of the lips. ii) Dental (D)-type; the closure section and the five sections directly behind it expand in such a way that a relatively smooth connection

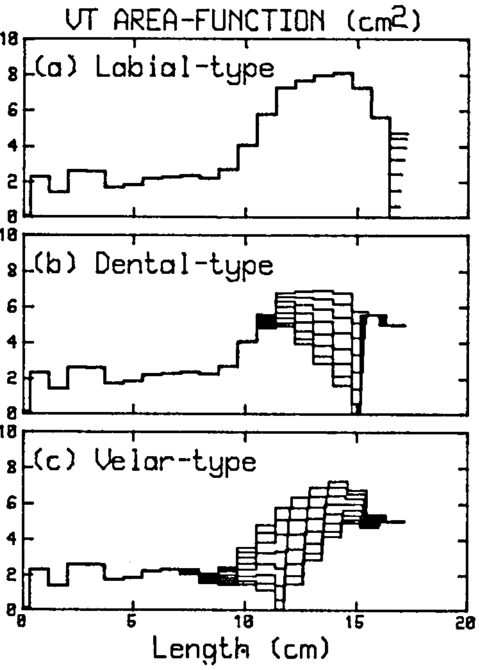


Fig. 1 Time-varying area functions for the three different closure types, sampled at every 20 ms following the release. For all cases, the target vowel is [a].

between the closure section and cavity behind (back-cavity) is maintained through release, as shown at Fig. 1b. D-type represents the dental gesture by the tongue apex and blade. iii) Velar (V)-type; The directly connected sections in both back and front cavity expands with the closure section, as presented at Fig. 1c. V-type is intended for the velar (or palatal) gesture involving the dorsum.

The way of the expansion of the closure section is specified by an exponential rise function. The cross-areas of the sections in the front and back of the occlusion also expand exponentially, but its onset rise is smooth without discontinuity. The smooth rise was necessary to prevent multiple excitation at vicinity of the occlusion. Resultant time-varying area function is fed to an acoustic VT simulator [2] for synthesis or for VT transfer calculations.

IDENTIFICATION TEST

In preliminary experiments, the three stops [p, t, and k] were synthesized with a reasonable quality by appropriately varying the value of the two articulatory parameters (the "position" along VT length and the "shape", L-, D-, or V-type), while the other parameters, such as VOT (=25 ms), glottal and release dynamics, were kept to fixed values. CV-tokens, therefore, were prepared by varying systematically the position, n, from 1 to 9, where the closure is located at the n-th VT section from the lips (n=1), and by varying the shape, for the three different target vowels, [i, a, and u]. The stimuli were randomized with five repetitions of each token for each closure shape.

Three experienced and three naive listeners participated in the test. Each listener was asked to identify the consonant, [either p, t, or k], and then type in the corresponding key on a computer key-boad. The next token was presented to the listener 1 sec after the response. The listeners were provided also a repeat request option upon which the same token is repeatedly presented. The total number of the repeats for each token was used in the interpretation of data as a measure indicating a quality of consonants.

The identification score for [p, t, and k], as a function of the position, for the target vowel [a] is presented for L-type at Fig. 2a, for D-type at Fig. 2b, and for V-type at Fig. 2c. The number of repeat requests for each token is also plotted in Fig. 2 by the dashed lines. As expected, the position of the occlusion is an essential factor in the production of the stops. Notice, however, that [t] cannot be produced by L-type (at Fig. 2a). The score for this consonant is only 37 % at best. For the same position (n=4), with D-type (at Fig. 2b) or with V-type (at Fig. 2c), the score reaches 100 %. For the velar [k], the score with L-type (at Fig. 2a) is relatively high, about 80 %. The number of repeat requests, however, is great, about 10 times, indicating an uncertain quality of the sounds as [k]. On the contrary, with V-type (at Fig. 2c), the identification function for [k], in fact for all three consonants, exhibits an ideal "categorical" response. Notice that the number of repeats, shown by the dashed line, increases only at the phonetic boun-

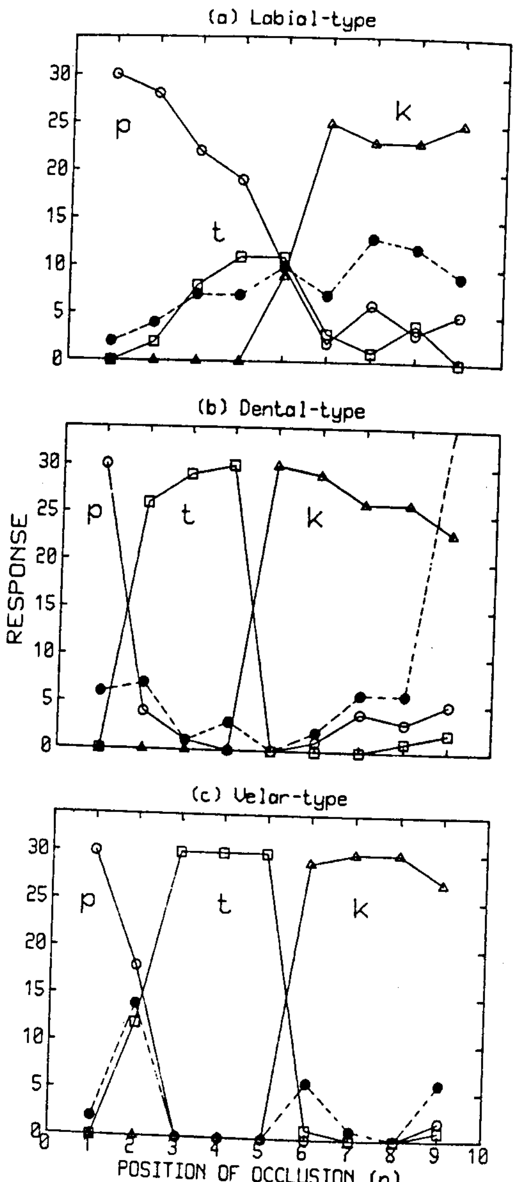


Fig. 2 The listener's responses as a function of the positions for the three different closure types. The full score, i.e. 100 %, corresponds to 30 on the ordinates (six listeners times five repetitions). The dashed lines indicate the total number of repeat requests by the six listeners for each token.

daries or at the extreme position (i.e., n=9).

For the other two target vowels, [i and u], the responses were similar to the case of [a] described just above, except labial-dental contrast in [i]-context. The consonant [p] with a high score (90 %) was produced only with L-type, while [t] with score 100% was only with D-type, for both at the position n=1.

It became clear that not only the position, but also the shape of the occlusion is an essential factor for the consonants to be identified correctly. The acoustic manifestations of these two articulatory parameters, therefore, must be relevant to the listener's processing.

The concomitant acoustic effects of the position and the shape of the occlusion are manifested in various form through out from the burst onset to the vowel transition. We shall attempt to sort out the acoustic characteristics that are consistent with the listener's reponses.

Due to the fact that during release, the aperture of the closure section is relativly small and the VT excitation sources are located at the exit of the closure, the acoustic characteristics related to only the front cavity appear on the burst [3]. The acoustics of the front-cavity is specified by the length (and thus by the position) and by how the cross-sectional area varies along its length. The listener's responses indicated that the dominant parameter for the velar, [k], is the position. Then, the spectral attribute of burst signaling the velar must be related to the front cavity, more specifically, the resonances of that cavity.

From the identification test, it is the shape that is more critical for the labial-dental distinction. If the aperture of the occlusion is, say, greater than 0.2 cm^2, the coupling effects appear on the transfer function and thus on the spectrum, regardless of the shape. When the aperture after release is still less than , say, 0.2 cm^2, the difference in the shape can manifest in the degree of the acoustic coupling between the front and back cavity. For a CV-syllable with L-type, due to the strong area-function discontinuity at the inlet of the occlusion (see Fig. 1a), the coupling is minimal. For the same CV, but with D- (or V-) type (see Fig. 1b), a smooth connection of the back-cavity to the occlusion is maintained from the onset through release. The coupling, therefore, is considerably enhanced in comparison with L-type.

In our synthesis, the rate of the area expansion at release was fixed to 20 cm^2/s. The presence or absence of the coupling effect, signaling the shape (L-type or D-type), therefore, can appear on the spectrum of the burst within 10 ms following the release. In other words, if the burst spectrum contains a series of peak and dip pairs (corresponding to that of pole-zero pairs in the transfer function) which is the indication of the coupling, then the shape is D-type or V-type. The absence of such spectral attribute implies the L-type.

The burst spectrum, therefore, can contain rich information to determine the position and the shape, which is essentially vowel-context independent. It is noted however that the effects of the two parameters can appear on the vowel transition, which are context dependent. In the following sections we shall examined a qualitative correspondence between the consonant identity (i.e., place) and spectral attributes of burst, and vowel transition for particular cases.

SPECTRAL ATTRIBUTES OF BURST AND PLACE

Attribute Pole-Zero for dental [t]

Burst spectra of the synthetic CV's, where the shape is D-type, and the target [i], are shown in Fig. 3. The position is varied from n=1 (closure at the lips) presented at the top in Fig. 3, to

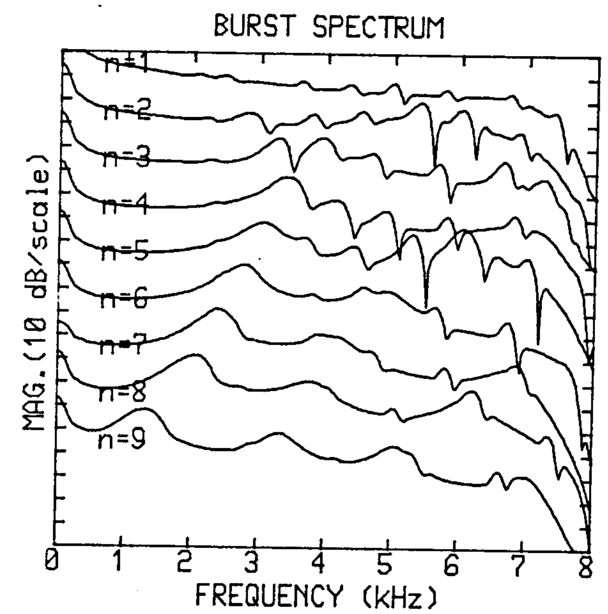


Fig. 3 **Burst onset spectra for nine different positions, from the lips (n=1) to the posterior extreme (n=9). The burst signals were synthesized with the target vowel [i] and the D-type closure.**

n=9 (at the posterior extreme) presented at the bottom. For the calculations, an 8 ms half Hamming window was aligned to each release onset.

For the position n=1 shown in Fig. 3, the front cavity is absent, resulting in a "falling" spectrum, which might imply, therefore, the labial [p]. Indeed, Blumstein and Stevens [4] has proposed the "diffuse-falling" gross shape of the burst spectrum as an invariant property for labials. The identification function indicated, however, that this token scored 100 % as dental [t]. Notice that the spectrum (n=1) exhibits a series of peak-dip pairs, which is a typical signature of the presence of the coupling, and thus of D-type shape. Let us call this kind of spectral characteristics attribute "PZ". The identical CV token, except L-type instead of D-type, scored 90 % as [p]. In this case, the attribute-PZ was absent. It is then stated that the shape is more critical than the position in labial-dental distinction, at least, in this particular case.

When the position is located at a slightly posterior position in the VT, i.e., n=2 or 3, a broad peak, BP, (at around 5.5 kHz for n=2 in Fig. 3) appears as the resultant effect of the front-cavity resonance. Thus BP might be considered as the attribute signaling dental, since the position is appropriate for the dental [t]. The presence of BP at a high frequency results in a rising mid-frequency spectrum, which may correspond to the invariant property "diffuse-rising" for dentals [4]. In our data, however, attribute PZ predicted more consistently the listener's identification of the dentals with a high score. An explanation for this will be described latter.

Attribute Prominent-Peak for velar [k]

At the position n=4 or greater, the length of the front cavity becomes relatively long, and the resonance frequencies shift toward lower frequencies and can exhibit a prominent peak, PP, as seen in Fig. 3. The presence of PP was common to the burst spectrum identified as [k] with a high score. The invariant property "compact" for velars [4] may correspond to the attribute PP. The presence of PP means a long front-cavity, and then an appropriate position as velar.

It should be mentioned that the CV tokens corresponding to n=7, 8, and 9 were considered as , at best, an ambiguous stop by the listeners and not velar, even though the skewed but prominent peak is present at low frequencies in each burst spectrum. The skewed peak is due to the rapid shift of the corresponding free-pole toward higher frequencies as the aperture of the constriction expands following release. When the shape, V-type, is employed for the otherwise identical token, the prominent peak is shifted toward a high frequency by more than 1 kHz, and indicated a less skewed peak. This is, of cause, due to the effects of the narrowed front-cavity toward the occlusion. Except for the extreme position (n=9) the corresponding tokens scored 100% as [k]. An inspection of the spectrograms shows that the shape, V-type, places the burst prominence at a "right" frequency in relative to the F-pattern in the vowel transition. In the case of the tokens with a high score, or in particular, of natural tokens, an appropriate position for velar implies the appropriate shape, and then the prominence at the right frequency. Consequentry, the presence of the attribute PP alone would suffice to specify velar.

Rules for predicting place

The prediction of place for unvoiced stops from the burst spectrum became evident. First, the presence of the attribute PP signals velar [k]. If absent, it means the position is for dental or labial. Therefore, if the attribute PZ is present, the consonant is dental [t], since it must be produced with the shape, D-type. If absent, then it is the labial [p].

THE SHAPE AND FORMANT TRANSITIONS

We shall concentrate our attention to the question why the shape is critical in the contrast labial vs. dental. A pertinent example was found in tokens with the target vowel [i]. As mentioned before, for the same position, n=1, token with L-type is identified as [p], wheares that with D-type as [t]. The spectrograms of the two tokens are shown in Fig. 4. Observe that F3 (the third formant) for the L-type at Fig. 4a is clearly rising, while F3 for the D-type at Fig. 4b is slightly but lowering. It has been demonstrated for voiced stops [5] that the F3 transition plays an important role in their identification, that is consistent with the effect of the shape described here. A similar effect of the shape on F2 transition was observed for the target vowel [a], (i.e., [pa vs. ta]).

It can be stated, then, that the influences of the shapes, L- or D-type, are manifested coherently on the burst and on the F-transitions, assuring a robust identification. This is the reason, probably, why the attribute PZ is favored over BP.

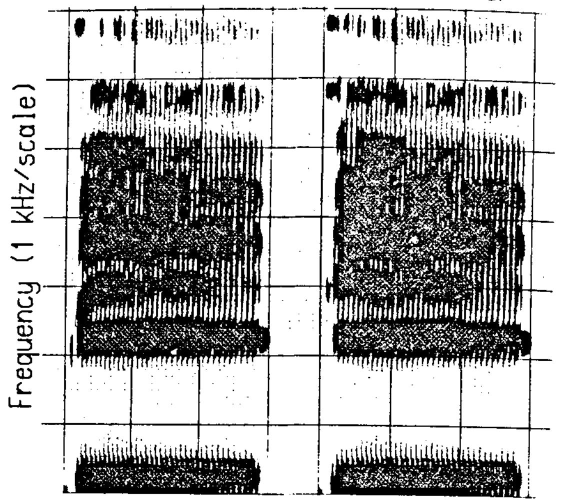


Fig. 4 Spectrograms of two CV syllables where C is identified as [p] at (a), and as [t] at (b). For both cases, the closure position is at the lips (n=1), and the target vowel is [i].

CONCLUDING REMARKS

In informal listening, it was often found that the identification of the stops upon signals corresponding to the burst alone or to the vowel part alone was difficult or impossible. When they had been assembled forming normal CV-tokens, however, the consonant was easily identified. From such experience, it is tempted to speculate that a suppression and/or an enhancening mechanism over the distributed cues are operating in the listener's processing. Coherent attributes found in both burst and vowel transition enhance each other. On the contrary, inconsistent attributes are suppressed. Such mechanisms may explain the listener identification functions in more comprehensive way.

REFERENCES

[1] S. Maeda, "Une source d'excitation coherente dans les occlusives," 15e JEP GALF, Paris, 43-46, 1985.
[2] S. Maeda, "A digital simulation method of the vocal-tract system," Speech Communication, 1, 199-229, 1982.
[3] G. Kuhn, "Stop consonant place perception with single-formant stimuli: Evidence for the role of the front-cavity resonance," J.Acous.Soc. Am., 65(3), 774-788, 1979.
[4] S. Blumstein and K. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J.Acous.Soc.Am., 66(4), 1001-1017, 1979.
[5] K. Harris, H. Hoffman, A. Liberman, P. Delattre and F. Cooper, "Effect of third-formant transitions on the perception of the voiced stop consonants," J.Acous.Soc.Am., 30(2), 122-126, 1958.

MODELING THE ACOUSTIC CHARACTERISTICS OF CHILDREN'S SPEECH: FUNDAMENTAL FREQUENCY

Corine Bickley

Department of Electrical Engineering and Computer Science
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

This paper presents a model of the vibration of child-sized vocal folds. The model reflects the anatomical differences between children and adults in laryngeal structure. A scale factor, or ratio of child to adult fundamental frequency, reflects these differences. For a one-year-old child, a scale factor of 4 is derived from the model. Values of fundamental frequency are predicted and are shown to be in agreement with values measured for young children.

INTRODUCTION

Children begin to produce speech-like sounds at a very early age. Normal children communicate with speech and language skills which approximate those of adults by the age of two or three years. The changes in sound production which take place during the first few years of a child's life result from changes in the child's anatomy and in motor-control and cognitive abilities. Each of these factors constrains the sounds produced by a child. Some of the acoustic characteristics of children's sounds are a direct consequence of the size and configuration of the structures involved in speech production: the lungs, the larynx, the vocal tract. Other characteristics may be influenced most by the motor-control skills of a young child. The cognitive ability to form and manipulate mental representations of words also has a significant influence on the sound sequences produced by a child.

One aspect of an examination of children's speech is modeling the acoustic characteristics of sounds. Models which predict the acoustic characteristics of adult speech abound in the literature, but only a few models of children's vocal systems have been proposed. An approach to predicting acoustic characteristics of children's speech is uniform scaling of all vocal-tract dimensions. This simple model fails to predict spectra which are in close agreement with measured spectra of children's utterances [13]. However, vocal-tract models which incorporate more detailed anatomical constraints, such as Goldstein's [4], generate formant frequencies appropriate for children.

Analyzing the source characteristics of children's speech remains problematic. A high fundamental frequency ($F0$) is a hallmark of children's speech. The mechanisms by which children produce and control these high fundamental frequencies are not well understood. Various models, including the vibrating string and spring-mass models, have been proposed to account for the fundamental frequencies used in speech. The vibrating string model predicts the general trend of higher fundamental frequencies of children's speech than adults', due to the differences in the lengths of children's and adult's vocal folds. The spring-mass models have been successful in predicting values for the fundamental frequencies of adult speech and for airflow through the glottis during vocal-fold vibration. Difficulties arise, though, in using these models to predict an appropriate ratio of a child's $F0$ to an adult's or in predicting reasonable values for children's $F0$ as a function of anatomical measurements.

BACKGROUND

The various theories of vocal-fold vibration indicate that the frequency of vibration and the shape of the airflow waveform depend on properties of the vocal folds, including the dimensions of length, thickness, and height (vertical thickness) and the Young's modulus and effective mass of the tissue.

Measurements

Measurements have been made of length and mass of the vocal folds, the thickness of the mucosa of the folds, and the stiffness of vocal-fold tissue. The length of the vocal folds has been measured for newborns, children and adults. Hirano et al. [5] report measurements of vocal-fold length, including both the membranous and cartilaginous portions, for males of various ages. Lengths for children and adults reported by Gedgoud [3], Negus [12], and Kahane [7] are summarized by Goldstein [4]. Several values are available for newborns and adults; relatively few are reported for children between the ages of one and seven years. Hirano et al. report an average length of approximately 3 mm for one-year-old children, or approximately one-sixth as long as the vocal folds of adult males.

The thickness of the mucosa of the vocal folds has been measured by Hirano and his colleagues for newborns, children and adults. The vocal fold thickens somewhat with age, but the change in thickness is not at great as the change in length. No direct measurements of vocal-fold height are reported. We assume that the change in height is comparable to the change in thickness.

It appears reasonable to assume that the vibrating mass of the vocal folds is proportional to the combined mass of the thyroarytenoid and lateral cricoarytenoid muscles and the vocal ligament. Kahane and Kahn [8] report the mass of the vocal fold muscles: 0.87 g for adults and 0.08 g for infants. Kaneko and his colleagues[9] estimated an effective mass of 0.14 g for adult vocal folds. Based on these values for adults and infants, we calculated effective vocal-fold masses for one- and two-year-old children of 0.02 and 0.03 g, respectively.

Vocal-fold stiffness has been measured for adult humans and for young and old dogs. Kaneko and his colleagues report an effective stiffness of 7.4×10^4 dynes/cm for the vocal folds of adult humans. Measurements of stress/strain relationships for vocal-fold tissue of young dogs and adult dogs were performed by Perlman and Titze [14]. They found that the vocal-fold tissue of young dogs is stiffer than the tissue of adult dogs. The vocal-fold stiffness K can be determined from measurements of Young's modulus and dimensions. From Perlman and Titze's graphs of stress vs. strain, we estimated a ratio of Young's moduli of young to old tissue of 1.3. Using this ratio, the stiffness reported by Kaneko et al. and vocal-fold dimensions, we computed a value of 2.1×10^4 dynes/cm for the stiffness of young vocal folds. This value is consistent the range of transverse moduli reported by Kakita et al.[10].

Models

Vibrating string and spring mass models have been proposed to describe vocal-fold vibration. For each of these models, the fundamental frequency of vibration of the vocal folds can be determined. The vibrating string model is a one-dimensional model whose parameters are vocal-fold length and tension. Various spring-mass models have been proposed (for example, [6]) which model the vocal folds in terms of lumped elements representing the mass, stiffness and losses of the vocal-fold structure. In order to predict the fundamental frequency of the vocal folds, only the effective mass and stiffness of the model are needed.

A scale factor, or ratio of child to adult male fundamental frequency, reflects the differences in anatomical parameters between children and adults. For the vibrating string model, the scale factor SF_{string} depends on vocal-fold length (L) and tension (T):

$$SF_{string} = \frac{F0_c}{F0_a} = \frac{L_a}{L_c}\sqrt{\frac{T_c}{T_a}} \quad . \tag{1}$$

The subscripts $_a$ and $_c$ refer to adult and child values, respectively. Assuming that the tensions T_c and T_a of child and adult vocal folds are approximately the same, we find that $SF_{string} \approx 6$ for a one-year-old child.

The scale factor for the fundamental frequency predicted by a spring-mass model is

$$SF_{spring-mass} = \sqrt{\frac{K_c}{K_a}\frac{M_a}{M_c}} \quad , \tag{2}$$

where K is the stiffness of the vocal-fold tissue and M represents the effective mass of the vibrating vocal fold. Solving for K in terms of the Young's modulus E and the dimensions of the vocal folds gives

$$SF_{spring-mass} = \frac{F0_c}{F0_a} = \sqrt{\frac{E_c}{E_a}\frac{h_c L_c}{b_c}\frac{b_a}{h_a L_a}\frac{M_a}{M_c}} \quad , \tag{3}$$

where h and b are the vocal-fold height and thickness, respectively. Assuming the same ratio of child to adult value for both cross dimensions h and b, the scale factor for the spring-mass model reduces to

$$SF_{spring-mass} = \sqrt{\frac{E_c}{E_a}\frac{L_c}{L_a}\frac{M_a}{M_c}} \quad . \tag{4}$$

For the values listed above, $SF_{spring-mass} \approx 1.3$ for the fundamental frequency of a one-year-old child compared to an adult.

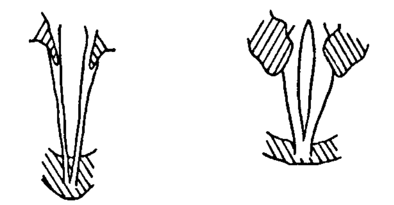
Both the vibrating string and spring-mass models predict that the $F0$ of a child's speech is greater than the $F0$ of an adult's speech. Neither prediction, however, gives a ratio which is in good agreement with the values of $F0$ of children reported by various researchers. Typical values of $F0$ for one- to two-year-old children are in the range of 300 – 500 Hz, or 3 – 4 times the $F0$'s reported for adult males.

THEORY

The vibrating string and spring-mass models capture important aspects of vocal-fold vibration, but fail to adequately model some aspects of the vocal-fold anatomy. For instance, the vibrating string model does not take into account the effect of the cross dimensions of the vocal folds on the stiffness of the structure. Another shortcoming of this model concerns the boundary conditions. The vibrating string model allows discontinuities in slope at the juncture of the cartilages and the vocal-fold tissue. The spring-mass model allows for discontinuities in both position and slope at the endpoints of the vocal folds. The specification of boundary conditions is important in analyses of the vibration of children's vocal folds; children's vocal folds are relatively shorter and thicker than adults', as shown in Fig. 1a. The attachments of the vocal folds to the arytenoid and thyroid cartilages can be expected to play a significant role in the vibration of children's vocal folds.

A model of vocal-fold vibration which reflects the anatomical structure of children's vocal folds is the bending beam model.

(a)



(b)

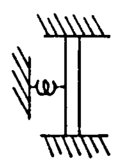


Figure 1: (a) Adult's and child's vocal-fold structures (not drawn to scale) (adapted from Bosma, 1986); (b) Bending beam model of vocal folds.

This model has been useful in predicting the vibratory motion of relatively stiff structures which are attached rigidly at their ends and vibrate a small amount in the transverse direction [15]. The traditional bending beam model can be augmented by the addition of a distributed stiffness along one side. Figure 1b shows a bending beam which is fixed at both ends and which is coupled to material on one side by means of a spring. The fixed ends model the attachment of the vocal fold to the arytenoid and thyroid cartilages. The spring models the lateral stiffness of the vocal-fold tissue.

The equation for transverse motion of the vocal-fold model shown in Fig. 1b is

$$\left(\frac{Eb^2}{12\rho}\right)\frac{d^4\xi}{dx^4} - \left(\omega^2 - \frac{K}{\rho L b h}\right)\xi = 0 \quad , \tag{5}$$

where E represents the Young's modulus of the vocal fold, K models the stiffness of the vocal-fold tissue, b and h are the thickness and height of the vocal fold, ρ is the density of the tissue, and ξ is the transverse displacement of the fold. Four boundary conditions are imposed: continuity of displacement and of slope at both ends of the vocal fold. A solution which is a linear combination of trigonometric and hyperbolic functions is assumed. Application of the boundary conditions results in

$$\cos\beta - \frac{1}{\cosh\beta} = 0 \quad . \tag{6}$$

The variable β takes on discrete values which are found by graphical solution; the lowest non-zero value of β is approximately 4.73.

The values of ω for which equation (5) has a solution are given by

$$\omega^2 = \frac{\beta^4}{12}\frac{Eb^3h}{L^3M} + \frac{K}{M} \quad . \tag{7}$$

The first term is the square of the natural frequency of the beam model of the vocal fold assuming no lateral stiffness, and is called ω_b^2. The second term, ω_s^2, is the square of the frequency of the spring-mass model of the vocal fold. The solution of the general equation of motion shows the combined contributions of the beam and the spring character of the vocal fold structure:

$$\omega = \sqrt{\omega_b^2 + \omega_s^2} \quad . \tag{8}$$

Numerical values for ω can be found by substitution of the dimensions and tissue properties of the vocal folds.

If boundary conditions of position continuity and no stress at the endpoints are assumed (instead of continuity of position and slope), the solution of equation (5) reduces to the solution of the equation of motion for a vibrating string.

RESULTS AND DISCUSSION

The frequency of vibration, ω, of the vocal fold is a combination of the terms ω_b and ω_s. The first term of equation (7) represents the frequency of vibration due to the characteristics of the beam, where

$$\omega_b = \sqrt{\frac{\beta^4}{12}\frac{Eb^3h}{L^3M}} \quad . \tag{9}$$

For small L, as in the case of a child's vocal folds, this term dominates the expression for ω, and $\omega \approx \omega_b$. The vibration of a child's vocal folds appears to be most like that of a bending beam. For a child-sized vocal fold with length 0.35 cm, height and thickness 0.23 cm, and mass and stiffness as above, we find $\omega_b = 2810$ and $\omega_s = 1000$. The child's fundamental frequency is thus 470 Hz.

In the adult case, or for large L, the ω_s term dominates, or $\omega \approx \omega_s$, where

$$\omega_s = \sqrt{\frac{K}{M}} \quad . \tag{10}$$

The vibration of adult-sized vocal folds is similar to the vibration of a mass coupled to a spring. For an adult whose vocal folds are of length 1.7 cm, height and thickness 0.27 cm, and mass and stiffness as above, $\omega_b = 160$ while $\omega_s = 760$. The corresponding fundamental frequency is 120 Hz.

Returning to our discussion of scale factors, we calculate a scale factor relating the $F0$ of the bending beam model (appropriate for a child's vocal folds) to the $F0$ of a spring-mass model (for an adult).

$$SF = \sqrt{\frac{\beta^4}{12}\frac{E_c b_c^2}{L_c^4}\frac{b_a^2}{E_a}} \quad . \tag{11}$$

For values listed above, $SF \approx 4$.

Measurements of $F0$ of comfort-state vocalizations of young children have been reported by several researchers. Keating and Buhr [11] report $F0$ measurements for children of ages 8 months to approximately 3 years. In a study of the acoustic characteristics of vowels produced by young children of ages one and one-half to two and one-half years, we found average values of $F0$ ranging between 350 and 400 Hz[1]. These values as well as those of Keating and Buhr are shown in Fig. 2. Overlaid on these values are predicted values at ages one, two and three years. It can be seen that the predictions of the bending beam model closely approximate the data for young children.

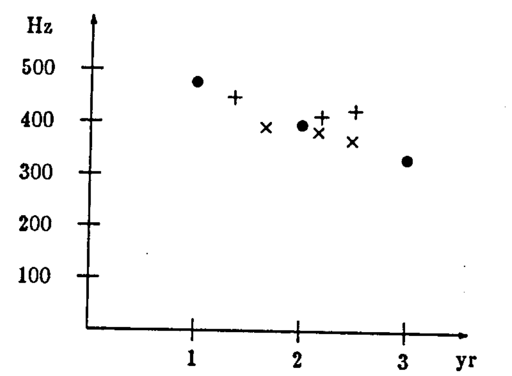


Figure 2: Predicted and measured values of $F0$. Predicted values are shown by filled circles. Averages of values reported by Keating and Buhr are shown by +'s; those of Bickley, by ×'s.

CONCLUSION

A model of vocal-fold vibration has been presented for which the expression for the fundamental frequency consists of two terms. The bending beam term depends on tissue characteristics and the connections at the ends of the vocal folds; the spring-mass term depends of the bulk characteristics of the folds. For young children, the bending beam term dominates; for adults, the spring-mass term determines the fundamental frequency.

Acknowledgement

This work was supported in part by a Whitaker Health Sciences Fellowship and in part by a LeBel Fellowship.

REFERENCES

[1] Bickley, C.A. *Acoustic Evidence for the Development of Speech.* Cambridge, MA: unpublished Ph.D. dissertation, Massachusetts Institute of Technology, 1987.

[2] Bosma, J.F. *Anatomy of the Infant Head.* Baltimore: Johns Hopkins University Press, 1986.

[3] Gedgoud, V.A. Anatomical peculiarities of the respiratory organs in children. Translated by S. Pelvoy, 1957. St. Petersburg: thesis. Referenced in U.G. Goldstein (1980) *Articulatory Model for the Vocal Tracts of Growing Children.* Cambridge, MA: unpublished Ph.D. dissertation, 1900.

[4] Goldstein, U.G. *Articulatory Model for the Vocal Tracts of Growing Children.* Cambridge, MA: unpublished Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[5] Hirano, M., Kurita, S. and Nakashima, T. The structure of the vocal folds. In M. Hirano and K.N. Stevens (Eds.), *Vocal Fold Physiology.* Tokyo: University of Tokyo Press, 1980.

[6] Ishizaka, K. and Matsudaira, M. What makes the vocal cords vibrate. In Y. Kohasi (Ed.), *Reports of the 6^{th} International Congress on Acoustics.* Tokyo: Maruzen, 1968.

[7] Kahane, J.C. The developmental anatomy of the prepubertal and pubertal larynx. Pittsburgh: unpublished Ph.D. dissertation. Referenced in U.G. Goldstein (1980) *Articulatory Model for the Vocal Tracts of Growing Children.* Cambridge, MA: unpublished Ph.D. dissertation, 1975.

[8] Kahane, J.C. and Kahn, A.R. Weight measurements of infant and adult intrinsic laryngeal muscles. Folia phoniat. 36:129-133, 1984.

[9] Kaneko, T., Masuda, T., Shimade, A., Suzuki, H., Hayasaki, K. and Komatsu, K. Resonance characteristics of the human vocal fold in vivo and in vitro by an impulse excitation. In T. Baer, C. Sasaki and K. Harris (Eds.), *Laryngeal Function in Phonation and Respiration.* Boston: College-Hill Press, 1987.

[10] Kakita, Y., Hirano, M. and Ohmaru, K. Physical properties of the vocal fold tissue: measurements on excised larynges. In M. Hirano and K.N. Stevens (Eds.), *Vocal Fold Physiology.* Tokyo: University of Tokyo Press, 1980.

[11] Keating, P. and Buhr, R. Fundamental frequency in the speech of infants and children. J. Acoust. Soc. Am. **63(2)**:567-71, 1978.

[12] Negus, V.E. *The mechanics of the larynx.* London: Heinemann Medical Books. Referenced in U.G. Goldstein (1980) *Articulatory Model for the Vocal Tracts of Growing Children.* Cambridge, MA: unpublished Ph.D. dissertation, 1929.

[13] Nordström, P.-E. Attempts to simulate female and infant vocal tracts from male area functions. Speech Transmission Laboratory Quarterly Progress and Status Report, Royal Institute of Technology, Stockholm, Sweden. **2-3**:20-33, 1975.

[14] Perlman, A.L. and Titze, I.R. Measurements of viscoelastic properties in live tissue. In I.R. Titze and R.C. Scherer (Eds.), *Vocal Fold Physiology: Biomechanics, Acoustics and Phoniatory Control.* Denver: Denver Center for Performing Arts, 1983.

[15] Woodson, H.H. and Melcher, J.R. *Electromechanical Dynamics.* New York: John Wiley and Sons, 1968.

ACOUSTIC-MECHANICAL FEEDBACK IN VOCAL SOURCE-TRACT INTERACTION

U. K. LAINE
Acoustics Lab.
Helsinki Univ. of Technology
Otakaari 5 A
02150 Espoo, Finland

E. VILKMAN
Phoniatric Dept.
Tampere Univ. Central Hospital
33520 Tampere, Finland

ABSTRACT

A new method to investigate vocal source-tract interaction is introduced. The method is based on the usage of excised larynges connected to an artificial vocal tract. Measurements of one larynx with a somewhat special behavior are described and analysed in detail. The analysed case gives clear evidence that the resonances of the vocal tract may influence directly or indirectly the vocal fold vibrations.

INTRODUCTION

In recent linear models for vocal fold vibration the vibratory pattern of the folds, i.e. the glottal opening as a function of time, is assumed to be an independent phenomenon in the sense that the vocal tract resonator has no effect on the mechanical vibrations of the folds. Until now the source-tract interaction has mainly been studied on the level of acoustic impedances, where the glottal opening and the subglottal tubes form an acoustic load for the vocal tract. Thus some part of the energy is lost from the vocal tract during every open period of the glottis [1].

It is well known that the sound pressure level (SPL) in the vocal tract just above the glottis is about 120-130 dB during voiced sounds. This study was undertaken to determine if this pressure is able to produce changes in the vibratory pattern of the vocal folds by deforming the mucosa-cover of the folds or by means of some other mechanism. In other words: Is there any acoustic-mechanical feedback in the vocal source-tract interaction?

We used excised larynges in our study. This is a legitimate method, known for instance from the work of van den Berg and Tan (1959) [2]. The novel methodological aspect of this study is that we combined excised larynges with an artificial vocal tract. This method makes it possible to control the resonances of the tract in a known and repeatable way. Since an artificial vocal tract is used, the changes of its profile will affect the vocal folds *only acoustically*. Therefore, we are able to distinguish between the mechanical (i.e. movements of articulators transferred via tissues) and purely acoustical effects. In our method only the acoustic power can affect the vibratory pattern of the vocal folds.

However, the method of van den Berg and Tan has severe limitations. First of all, it is almost impossible to simulate the action of the thyroarytenoid muscle [2], [3], and second, the dead tissue does not permit accurate measurements of the vibratory pattern of the vocal folds over a longer period of time [4]. The first problem is not a serious one, as the body, i.e. the vocal muscle is not of great importance in pitch control of phonation [5]. The vulnerability of the cover (mucosa) of the vocal fold was pointed out already by van den Berg and Tan [2]. The second problem may be solved by limiting the duration of each phase of the experiment and performing an adequate number of repetitions and by stabilizing the arrangement of each phase.

This study was carried out at the Phoniatric Department of the Tampere University Central Hospital in cooperation with the Acoustics Laboratory at the Helsinki University of Technology. In the Phoniatric Department this study is part of a larger long-range project investigating questions in voice physiology. In the Acoustics Laboratory this study is part of a chain of studies dealing with the modelling of speech acoustics.

The experiments we made produced a bulk of material that needs to be studied in more detail. In this preliminary report we concentrate on one of the most interesting phenomena observed.

MATERIAL AND METHOD

The effects of acoustic-mechanical feedback on the vibrations of the vocal folds were examined in three fresh excised larynges taken from autopsies of males. In the dissection the vocal folds were left intact. The epiglottis and the ventricular folds were removed in order to get a better view of the vocal folds [2], [4]. After dissection the specimens were stored in 0.67% NaCl solution at a temperature of +4 °C for 1-2 days.

One of these larynges showed an exceptional high sensitivity to the variations of the supraglottal resonances and therefore it was chosen for closer analysis. It was obvious that acoustic-mechanical feedback in the vocal source-tract interaction should be seen most clearly in this case.

The experimental arrangements are shown in Fig. 1. For the experiment the cricoid cartilage was fixed in an air tight manner on an acrylic plate just above the hole for air intake. The supraglottal acrylic tube (length 17.5 cm, inner-diameter 2.9 cm, volume 115.6 cm^3) was attached to the thyroid cartilage and supported with a holder. An air-tight connection of the tube-thyroid cartilage junction was obtained by using plastic mass (Optosil®) and rubber sealant.

The glottal closure was obtained with two threads attached to each arytenoid cartilage. A constant force was used to pull each thread throughout the experiment. Phonation was elicited by a constant humidified and warmed (37°C) air flow which passed through the acrylic plate. The flow was measured using a flow meter (AGA). Under the acrylic plate was a sampler for condensation water. The sampler acted as the subglottal space [6]. On the side of the sampler there was an outlet for measurement of the subglottal pressure, which was recorded (Frökjaer-Jensen Manophone).

The acoustic load of the artificial vocal tract, i.e. the supraglottal tube, was varied by moving an acrylic cylindrical block in the tube. The position of the cylinder was visually monitored by using a centimeter scale drawn on the tube. The block was 8 cm in length and 2 cm in diameter. This choice was made so as to reserve free space for the cable of the photosensor (see Fig. 1). With this block we were able to vary the frequency of the first formant from 400 to 600 Hz.

The subglottal pressure varied between 10-20 cm H_2O. This is somewhat high for speech but still within physiological limits. The average pitch of this larynx was about 170-180 Hz, higher than in a normal male voice.

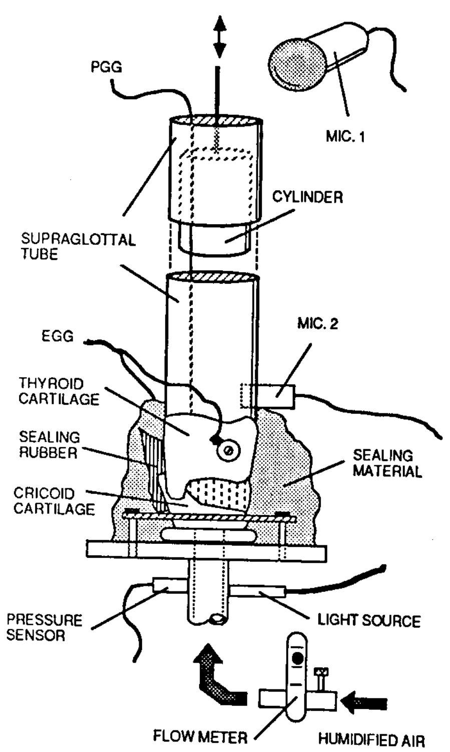


Fig. 1 *The experimental arrangement.*

Electrical signals describing the vocal fold vibrations were recorded by using a high-quality tape recorder (Tascam, acoustical signals only), a FM-type instrumentation recorder (Racal) and a digital PCM coder and recorder (Sony Digital Audio Processor F-1 and Portable Video Casette Recorder SL-F1E). One acoustic microphone (AKG C5657E) was placed close to the opening of the supraglottal tube and the other (B & K 4133) was air-tightly mounted into a hole on the side wall of the tube just above the glottal level. The electroglottographic signal (Frökjaer-Jensen EG 830) was obtained using small coin-shaped brass electrodes attached with a screw symmetrically to each side of the thyroid cartilage on the vocal fold level [4]. The photo-electric glottographic signal was obtained by introducing a light beam into the subglottal space through a window (Frökjaer-Jensen Photo-electric Glottograph). The light which passed through the glottis was detected by a photosensor placed in the supraglottal tube.

The vibratory pattern of the vocal folds was monitored using a laryngostroboscope (B & K Type 4914). The recorded signal samples were analysed at the Acoustics Laboratory of the Helsinki University of Technology using a PC-based (MacIntosh) ISA-system (Intelligent Speech Analyser®, Vocal Systems, Ltd).

RESULTS

One of the larynges showed a special behavior. Its vibrations were typically weak and sounded leaky (noisy) and aperiodic, somewhat creaky. Changing the flow or subglottal pressure did not improve its performance. Only when the block was put in the resonator did the vocal folds start to vibrate strongly with stable amplitudes and periods. When the block was placed in deeper, simulating a back vowel, the vibrations once again became weak and inconsistent. The vibrations were strong only when the block was about in the middle of the tube or in the front.

Fig. 2 illustrates how the subglottal pressure varies when the block is moved from a back vowel position out of the tube. Initially the vocal folds are not vibrating properly, the glottis is leaky and the subglottal pressure low. When the block is moved upwards a stronger vibration suddenly starts and the pressure increases indicating a better glottal closure. When the block is out of the tube the vibrations are again weak and the pressure low. This was a systematic and repeatable phenomenon achieved with this larynx. During this experiment the photosensor was removed to make the movements of the block free and to ensure that the possible movements of the sensor were not creating this phenomena.

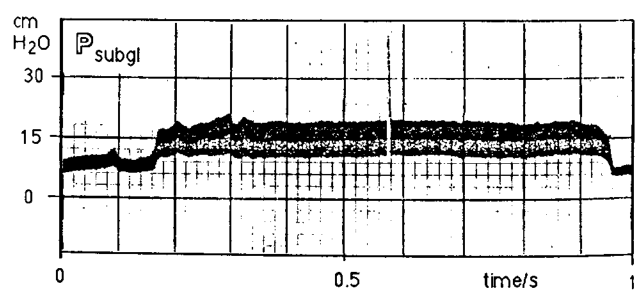


Fig. 2 *Change in the subglottal pressure due to the variation in the supraglottal impedance and in the vocal fold vibration.*

Fig. 3 shows in more detail how the stronger vibration begins and ends. In this figure the DC component has been removed.

Two spectra of the subglottal pressure are seen in Fig. 4. The upper part of the figure shows the signals where the stronger vibrations began and the lower part where they ended. One can note that the increase in the amplitude of this signal is mainly due to the increase of its first harmonic. The amplitude of the fundamental is not changed by much. The amplitudes of the second and third harmonics have also increased. When the block is moved out of the tube and the intensity of the subglottal pressure signal falls off, the changes in the harmonic structure are about the same but in the opposite direction. The intensity of the first three harmonics are affected the most. The strongly increased levels of the first harmonics will also indicate a better closure of the glottis.

The variations in the outcoming acoustic signal are seen in Fig. 5. The general trend is the same as in the earlier figure. The two first resonances of the tube are located at about 600 Hz and 1.3 kHz indicating that the block is in the back vowel position. When comparing the upper parts of Figs. 4 and 5 one can note that during strong vibration, i.e. better glottal closure and higher subglottal pressure (indicated by the white spectrum), the harmonic peaks in the region of the second formant are not seen in the subglottal pressure signal, whereas when the glottal closure is bad these peaks are clearly seen (black spectrum).

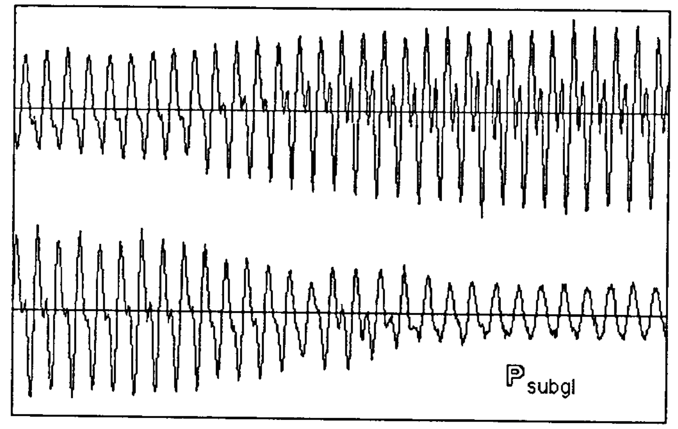


Fig. 3 *Subglottal pressure wave at the beginning and end of the stronger vocal fold vibration.*

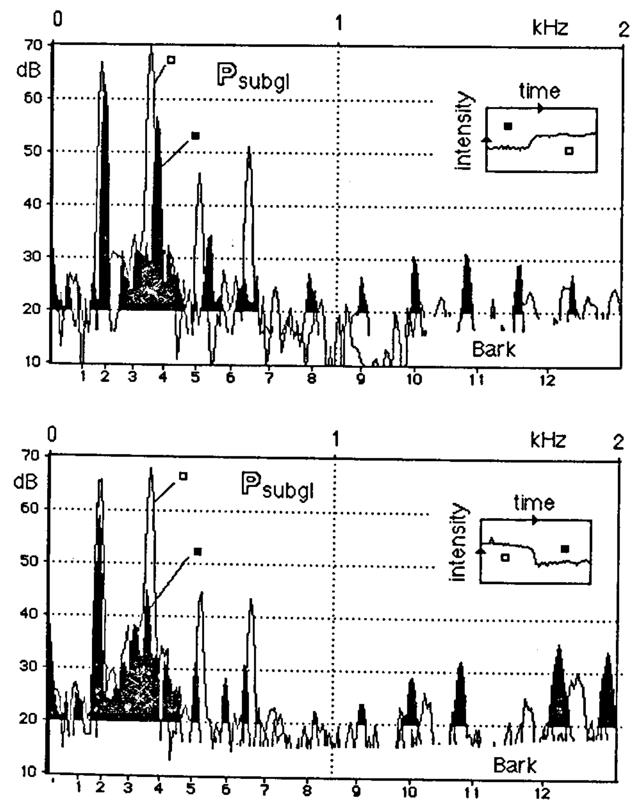


Fig. 4 *Change in the spectra of the subglottal pressure at the beginning and end of the stronger vocal fold vibration.*

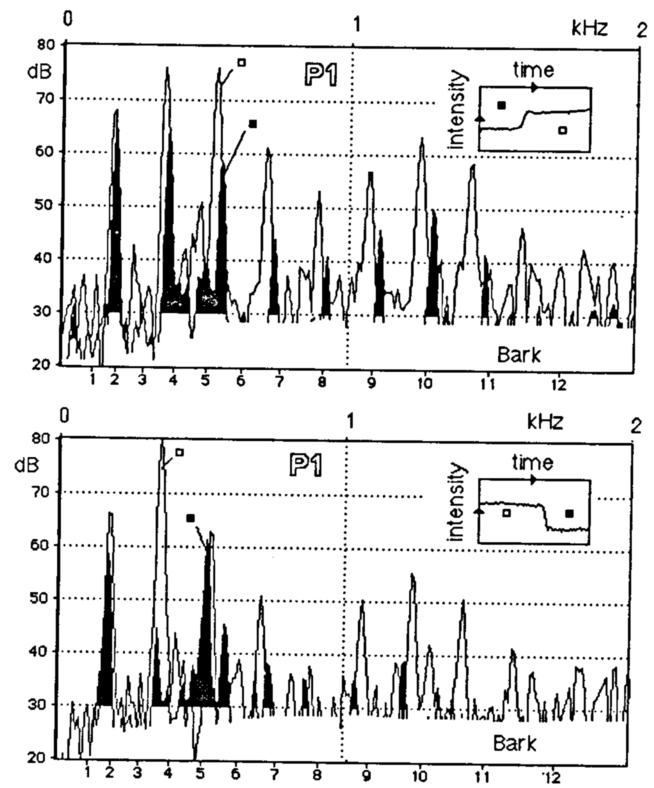


Fig. 5 *Change in the spectra of the outcoming acoustic signal (mic. 1).*

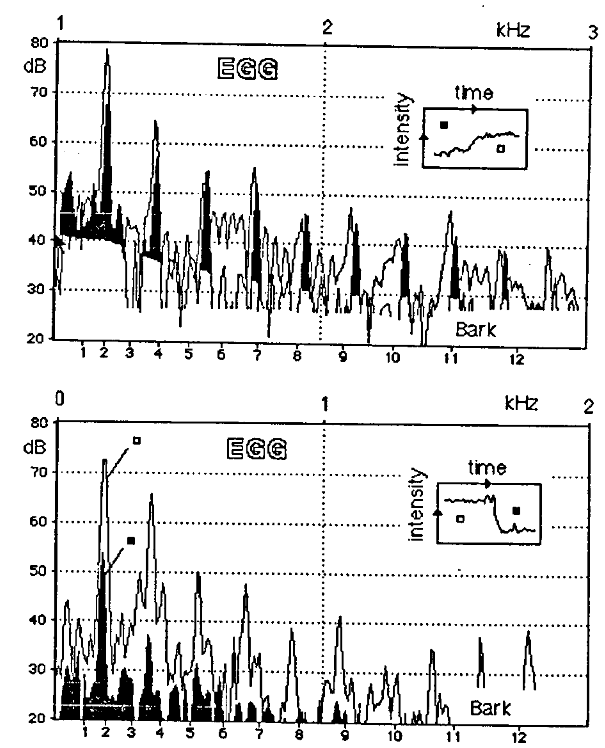


Fig. 6 *Change in the spectra of the EGG.*

Fig. 6 shows the corresponding variations in the spectra of the EGG signal. In the upper part of the figure (increasing intensity) the strongest amplitude change is seen at the peak of the fundamental frequency while the levels of the harmonics are not affected as much. When the intensity is decreasing the change is about the same over the whole spectrum.

Fig. 7 compares two EGG pulseforms normalized in amplitude and frequency taken from the first low intensity region and from the beginning of the high amplitude region. During weak oscillations the decreasing contact (opening) forms only about 20% of the pulse duration (pulseform 1). When the oscillation is strong the corresponding region is about 64% (pulseform 2). In this respect the EGG pulseform is changed radically even if the power spectra (Fig. 6 upper part) remains about the same. This indicates that *the phase relationships are changed.* The weak pulse (1) indicates that there are some types of acoustical forces coming from the tube resonator which are able to make the opening of the glottis faster and the closing slower. This breaks the vibratory pattern of the vocal folds and gives the voice a bad quality. In the opposite case the forces are in phase with the natural glottal oscillations and the voice quality is good.

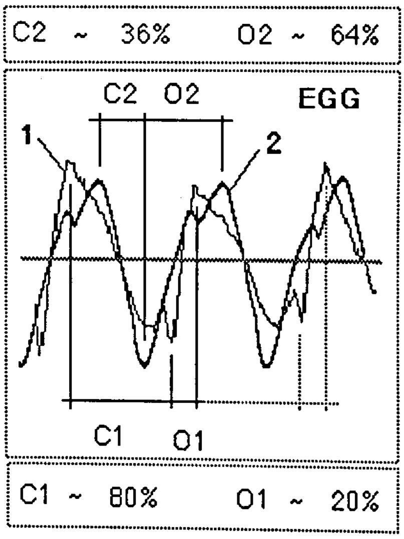


Fig. 7 *Normalized EGG pulseforms:*
1 ª at low intensity region — *C ª closing periods*
2 ª at high intensity region — *O ª opening periods*

DISCUSSION

Our new method of combining an excised larynx with an artificial vocal tract has given a clear indication that the vocal tract resonator is able to produce such a high acoustic energy above the vocal folds that their vibratory pattern may be radically affected. The acousto-mechanical phenomena we are investigating seems to be too complicated to be explained with present-day linear models. According to Mozer [7] the phase relation between the fundamental and the first harmonic may affect the vocal fold vibration. The pitch was relatively high in this case and we have estimated that the formant movement in question can make a phase change of about 90 degrees between the fundamental and the first harmonic. Therefore, the strong second harmonic when in optimal phase with the vocal fold vibrations may produce a better closure and otherwise may hinder the complete closure.

Titze [8] has also reported about this kind of interaction: "... it would appear that the vocal tract pressures reflected back to the glottis can assist in sustaining vocal fold vibrations."

Our results have confirmed this: the acoustic power in the vocal tract can assist or hinder the vibrations of the vocal folds. Does this feedback, which seems to be nonlinear, work directly on the mucosa cover of the folds or indirectly via the Bernoulli effect? This question still remains open.

ACKNOWLEDGEMENTS

The authors are grateful to Prof. A. Sonninen for the fruitful discussions during this work. This study is supported in part by the Academy of Finland.

REFERENCES

[1] **Ananthapadmanabha T., Fant G.,** Calculation of the True Glottal Flow and its Components, STL-QPSR 1, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, 1982, pp. 1-30.

[2] **van den Berg J. W., Tan T. S.,** Results of Experiments with Human Larynges. Pract. Oto-rhino-laryng., 1959, vol. 21, pp. 425-450.

[3] **Fukuda A., Saito S., Kitahara S., Isogai Y. et al.,** Vocal Fold Vibration in Excised Larynges Viewed with an X-ray Stroboscope and an Ultra-high Speed Camera. In: Bless D. M., Abbs J. H., eds.: *Vocal Fold Physiology.* San Diego, CA, College-Hill Press, 1983, pp. 238-252.

[4] **Lecluse F. L. E.,** *Elektroglottografie.* Dissertation, Rotterdam, Erasmus University, 1978.

[5] **Fujimura O.,** Physiological Function of the Larynx in Phonetic Control. In: Hollien H., Hollien P., eds.: *Current Issues in the Phonetic Sciences*, vol. 1., Amsterdam, John Banjamins B. V., 1979, pp. 129-164.

[6] **van den Berg J. W.,** Sound Production in Isolated Human Larynges. In: Bouhuys A., ed., *Sound Production in Man,* Ann. N.Y. Acad. Sci., 1968, vol. 155, pp. 18-27.

[7] **Mozer M.,** Artikulatorische Einflüsse auf die Stimmenreinheit, Sprache - Stimme - Gehör 9, Georg Thieme Verlag, Stuttgard, New York, 1985, pp. 117-120.

[8] **Titze I. R.,** Influences of Subglottal Resonance on the Primary Register Transition. In: van L. Lawrence, ed., *Transcripts of the 30th Symposium Care of the Professional Voice, Part I: Scientific Papers,* The Voice Foundation, N.Y., 1984, pp. 130-134.

THE METHOD FOR SOLVING INVERSE PROBLEM OF SPEECH PRODUCTION AND ARTICULATORY PORTRAY OF A SPEAKER

Yevgeni Vlasov, Natali Isayeva

Institute of Control Sciences
Academy of Sciences USSR
Moscow, USSR 117342

ABSTRACT

The accuracy of modern methods for determination of area function is not sufficient for practice. We present a numerical method for area function calculation with sufficient accuracy. Regulation continuum of this functions in finite region is called articulatory portray.

INTRODUCTION

Many year modelling of speech production processes has attracted investigators/1/, however its complexity up to the present does not lead to a wide introduction of such models into practice, despite the great efforts /2/ and intensively growing feasibilities. It has become clear, that the speech production processes are hierarchical and closely interacting /3/. In this case speech production model is expedient to be realized from bottom to top, using a lower level as a tool /4/. One of them is the articulatory model /5/, which synthesizes speech on the basis of solving the direct problem of speech production (vocal tract → acoustic) /6/.

The more accurate data of area functions have been obtained by G. Fant (1960) /7/ and up to now this work remains unique, because of its complexity. The LPC-method /8/ requires special measures (beforedistorting, etc.) for obtaining valid solutions. The tomography method /9/ enables us to determine the area in any section, however it requires multiple X-ray photographing of such sections along the axis for reconstructing only one area function. It was necessary to develop the method for an easier way of obtaining area functions without accuracy loss.

The paper presents the method for solving the inverse problem of speech production (acoustic → vocal tract), which allows the obtaining of "smooth" area functions and articulatory portray. The latter represents the region of permissible articulatory situations of a speaker. This method is based on the idea of "analysis from synthesis" and includes the algorithm /6/ with two-tier adaptive program complex (APC) /10/. The distinct features of the APC are automatical problem orientation to the class of the problems to be solved, supported by multidimensional optimization and associative information processing by a computer.

INPUT DATA AND ERRORS

Input data are easily-measured spectrum-time speech parameters: frequences, bandwidths and amplitudes of forments and also X-ray images of vocal tracts in segital flatness of three speakers: two men /7,11/ and one woman /12/.

The formant frequences have been determined by sonagraph, the errors were 3-7 per cent. In future calculations the frequences vector $F^* = (F_i^*, i = 1,k)$ will be the standard and errors vector $\varepsilon^* = (\varepsilon^*, i=1,k)$ will be final accuracy.

From X-ray images we used the samples $H = (H_i, i = 1,M)$ of a heigt function $h(x)$, $0 \leq x \leq \ell$, where ℓ is the vocal tract length. Samples and length errors are respectively equal to 7 and 3 per cent.

THE METHOD

Taking into account the difficulty of obtaining an X-ray images, the method is realized by two variants: with X-ray images and without them.

<u>The first variant</u> (with an X-ray image). The area function S(x) is represented as a product of the known height function h(x) on a desired width function

The finite articulatory region determines

$$D_q : \left(q^{min}(x) \leq q^{0}(x) \leq q^{max}(x), \quad 0 \leq x \leq \ell \right) \qquad (1)$$

where $q^0(x)$ – some initial approximation. Sampling of the all three function along the axis x in D_q gives respectively the vector of the lower boundary $W^{min} = (W_i^{min}, i = 1,N)$, initial control vector $W^0 = (W_i^0, i = 1,N)$ and the upper boundary vector $W^{max} = (W_i^{max}, i=1,N)$ as shown in Fig. 1

$$D_W : \left(W_i^{min} \leq W_i^0 \leq W_i^{max}, \quad i = 1,N \right). \qquad (2)$$

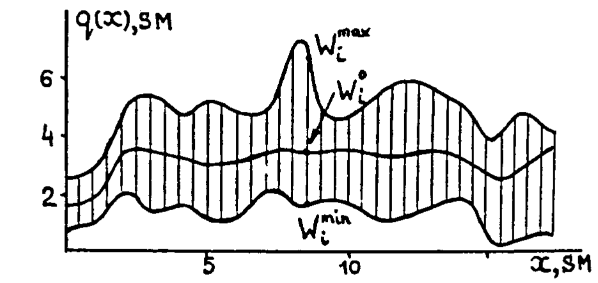


Fig. 1. Control vector W° is the samples of the function $q(x)$ in known boundaries (1)

Characteristics vector P, which characterizing the class of solved problems, includes F^*, ℓ , H:

$$D_p:(P=P_1,\dots,P_{K+M+1})=(F_1,\dots F_K,\ell,H_1,\dots H_M) \quad (3)$$

The mathematical statement of the speech production inverse problem has the following form:

$$L(F,F^*)\to min,$$

$$L(F,F^*)=\Big(\sum_{i:E_i\geq\varepsilon_i}(E_i-\varepsilon_i)^2\Big)^{1/2},\ E_i=\frac{|F_i-F_i^*|}{F_i^*},\ i=1,K,$$

$$F=G(S(x)),\ S(x)=h(x)q(x),\ 0\leq x\leq\ell, \quad (4)$$

$$h(x)=R(H),\ \ q(x)=R(W),$$

$$W\in(W_i^{min}\leq W_i^{\circ}\leq W_i^{max},\ i=1,N),$$

where L - functional, which depends on calculated frequencies F, G-an operator of the speech production direct problem, R - an operator transforming a given vector to a smooth function.
The method of solving problem is shown in Fig. 2 and consists of the following. Each problem is determined by concrete values of the vector P according to (3). Initial vector W° gives the random width function q(x), which determines the random area function S(x)=h(x)q(x). The operator G calculates F, which is compared with the standard F^* for determining L. The value L is analysed in the APC with the aim of optimizing the components for finding the minimum of L. When the final value L^* is achieved, the calculation process is finished and the decision vector (P,W^*,L^*) is stored in a computer memory. For a new problem P′ we take from the memory such an initial W° in the set of earlier solved problems, whose vector P is closer to P′.

The second variant (without an X-ray image). The region is determined as follows:

$$D_s:(S^{min}(x)\leq S(x)\leq S^{max}(x),0\leq x\leq\ell,\ \ell\in[\ell^{min},\ell^{max}]), \quad (1')$$

sampling of which, gives the region $D_W(2)$. The vector P includes the formant frequencies F^*, bandwidths ΔF^* and amplitudes A^*. The two latter vectors may be not available.

$$D_p:(P=P_1,\dots,P_{3K})=(F_i^*,\Delta F_i^*,A_i^*,\ i=1,K). \quad (3')$$

In (4) the area function is formed directly from the control vector:

$$S(x)=R(W),\ 0\leq x\leq\ell,\ \ell\in[\ell^{min},\ \ell^{max}] \quad (4')$$

In other aspects this varient does not change and is illustrated in Fig. 3 (compare with Fig. 2).

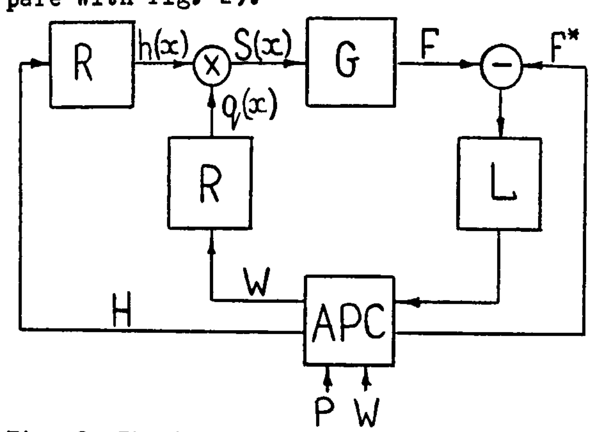


Fig. 2. Block-sheme shows the solution method for the first variant.

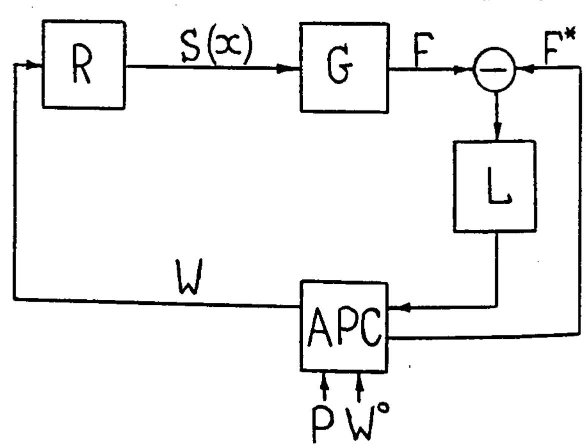


Fig. 3. Block-sheme shows the solution method for the second variant

REGULATION ALGORITHM

In general the inverse problems are mathematically noncorrect, i.e. they admit nonunique solution. Correctness of the considered problem is caused by constraining permissible solution region (1) and by the development of the special regulation algorithm.
This algorithm works along contour II in Fig. 4, while contour I denotes the functioning of the APC without it. At first

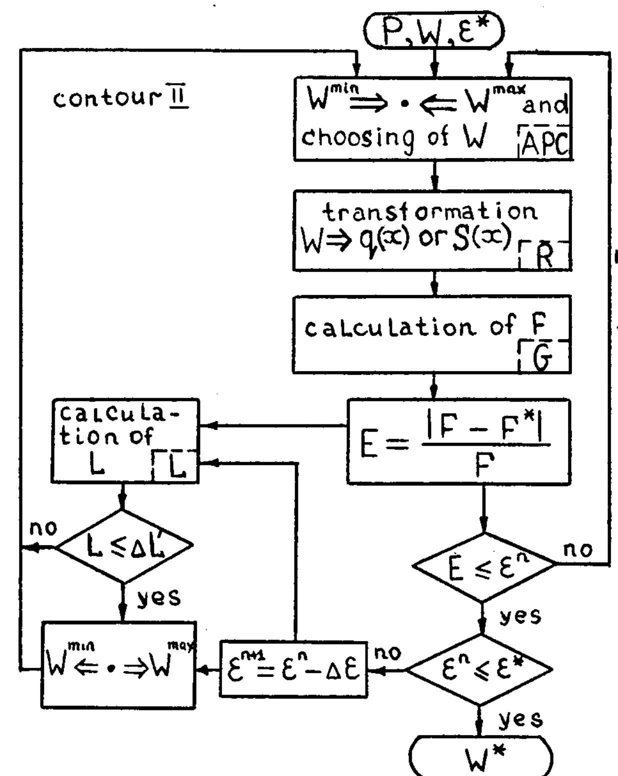


Fig. 4. The regulation algorithm

the rough stages value of accuracy ε° is assigned. For a given P and initial W° the error E is calculated and compared with ε° . In contour I the ε° is attained ($E\leq\varepsilon^{\circ}$), then contour II gives the next accuracy $\varepsilon^1=\varepsilon^{\circ}-\Delta\varepsilon$, etc. It should be noted that in contour I the boundaries (2) are approached: $W^{min}\Rightarrow\cdot\Leftarrow W^{max}$, while in contour II they are expanded: $W^{min}\Leftarrow\cdot\Rightarrow W^{max}$ up to initial boundaries (2). The algorithm is also adapted to the change of L: contour II is switched over if the reduction velocity ΔL less than the threshold $\Delta L'$.
Thus, the reaching final accuracy ε^* is divided into a sequential stages, each of which gains a stage accuracy E^n

$$\varepsilon^{\circ}\geq\varepsilon^1\geq\dots\geq\varepsilon^n\geq\dots\geq\varepsilon^*$$

$$W^{\circ}\Rightarrow W^1\Rightarrow\dots\Rightarrow W^n\Rightarrow\dots\Rightarrow W^* \quad (5)$$

In this case sequence $S^n(x)$ tends to optimal $S^*(x)$.

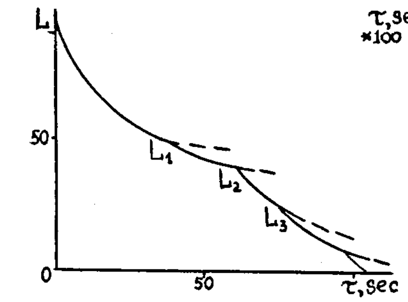
RESULTS

The proximity criterion of functions S(x) $S^*(x)$, simular to /8/, is the mean square deviation, normalized by a maximum

$$\sigma=\Big(\frac{1}{M}\sum_{i<1}(S_i-S_i^*)^2\Big)^{1/2}\Big/\max_{1\leq i\leq M}S_i^* . \quad (6)$$

Stability. The scatter of obtained solutions $S^*(x)$ under the variations of the initial approximation W° and boundaries W^{min} , W^{max} in (2) have been estimated. For similar phonems this scatter does not exceed 6.7 per cent with the deviation of control vectors from the initial values (Fig. 1) up to 120 per cent.

Convergence. The convergence to the accurate solution S (x) is guaranteed by the above regulation algorithm. Rejection of this algorithm leads to the interruption of convergence, as shown by the dotted line in Fig. 5. The continuous line shows the normal process of convergence: in points L_1 , L_2 , L_3 correction of stage accuracy ε^n takes place by contour II.



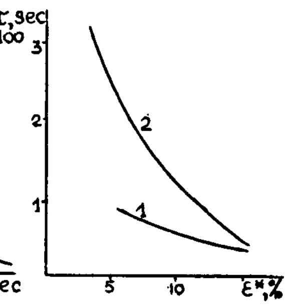


Fig. 5. Minimization of L shows the convergence of computer process.

Fig. 6. The illustration shows the benefit of regularization algorithm application.

Accuracy. Accuracy of $S^*(x)$ is estimated by (6) for the first speaker, since he has the exact area function /7/. Among the phonems the mean accuracy equals 8.3 per cent and it varies in the range of 4.1 - 12.9 per cent. With respect to the results /8/, where the LPC-method is used for the same speaker, the accuracy has increased by 2.7 per cent.

Computer time. The application of the requltion algorithm provides not only required accuracy, but acceleration of the computer processes as well. Fig. 6 shows that the final accuracy ε^* is move benificially obtained making use of this algorithm since the solution time with the algorithm application (curve 2) is many times reduced as opposed to the one without algorithm application (curve 1). The quantity of benefit is increased with the increasing of the final accuracy, from 1.6 times at ε^* = 10 per cent to 4.5 times at ε^* = 5 per cent. Additionaly, owing to optimal fitting of the algorithm parameters the computer time is reduced by 2.5 -30 times. Among the phonems the average computer time is equal to 84 sec and varies in the range of 4.5-148 sec.

Comparison of two variants. Different input data application (with / without X-ray images) leads to the average error of 9.4 per cent in solutions. The average computer time in the second variant is greater by 18.4 per cent than the one in the first variant. Hence, decreasing of apriori information should be compensated at the cost of increasing the computer time.

ARTICULATORY PORTRAY

Representation of the relationship of the solution $S(x)$ and input $h(x)$ as a functional dependence

$$S = S(h(x), x), \quad 0 \leqslant x \leqslant \ell \qquad (7)$$

in three-dimensional space (S, h, x) leads to a complex surface as shown in Fig. 7.

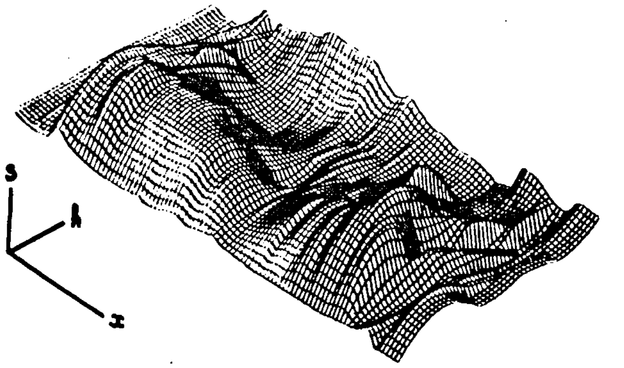


Fig. 7. Articulatory portray of a speaker

Such surface clearly presents the region of admissible articulatory situations of the speaker, therefore it is called an articulatory portray. Such protray vividly shows pharinx and oral (two saliences) and contraction of a larinx tube, a contraction caused by valum (pass between saliencea) and lips. The practical application of such a portray particularly consists in easy transformation of the flat X-ray image of vocal tract to area function with a given accuracy.

Comparison of portrays. The degree of individual distinction between the speakers is equal to 5.1 per cent for a diameter of a vocal tract and to 3.1 per cent for a longitudinal dimension, that may be interpreted as a value of articulation "quanta" /13/.

CONCLUSION

Vocal tract is usually approximated by a few cylinder sections, but in practice, a a more smooth area function is required. In proposed method there are no restrictions on a quantity of sections and computer time does not depend on the quantity of sections. The required accuracy, corresponding to input data errors, is also quaranteed. The method provides the reduction the compute time up to 4.5 sec, which equals one computer remembrance time. For articulatory synthesyzer /5/ smooth area function provides the improvement of the quality of a synthetic speech.

Computer time depends on the degree of APC-knowledge and the correct sequence of the problems to be solved. The better the APC is trained the shorter is the computer time. Due to the rational choice of problem sequences the training time decreases by 1.5-2.0 times with respect to random sequences.

In addition to it, this method may be used in speech analysis, medicine and in logophedia.

REFERENCES

/1/ J. Balazs, "In Memoriam Farkas Kempelen", Hungarian Papers in phonetics, ed. K. Bolla, No.13, 1984, p.11-21.

/2/ J. Allen, "A Perspective on Man-Machine Communication by Speech", Proc. IEEE 73 (11), 1985, p.1541-1550.

/3/ В.Н.Сорокин, "Теория речеобразования", Москва, Радио и связь, 1985.

/4/ P. Mermelstain, "Articulatory Model for the Study of Speech production", J. Acoust. Soc. Am. 50(4), 1973, p.1070-1082.

/5/ Е.В.Власов, "Акустический терминал для артикуляционного синтезатора речи", Тезисы I2-го Всесоюзного семинара по автоматическому распознаванию слуховых образов, Киев, Институт кибернетики АН УССР,1982,с.389-393.

/6/ Е.В.Власов,"Модификация метода Галеркина для расчета частотных параметров речевых сигналов", Проблемы построения систем понимания речи,Москва, Наука, 1980, с.I36-I42.

/7/ G. Fant, "Acoustic Theory of Speech Production", Gravenhage, Moution, 1960.

/8/ J.D. Markel, A.H. Gray, "Linear Prediction of Speech", N.Y., Springer-Verlag, 1976.

/9/ S.Kiritani, E. Takenaka, M. Sawashima, "Computer tomography of the vocal tract", Ann. Bull. Research Inst. Logopedics and Phoniatrics, Tokyo, No.12, 1978, p.1-4.

/10/ В.С.Широколава, Н.А.Исаева, "Двухъярусный обучающийся программный комплекс", Модели управления сложной программой, Москва, Институт проблем управления, I986, с.3-I0.

/11/ В.Н.Сорокин, "Механика движений языка", Описание и распознавание объектов в системах искуственного интеллекта, Москва, Наука, I980, с.42-7I.

/12/ K. Bolla, "A Phonetic Conspectus of Russian", Hungarian Papers in Phonetics, No.11, 1982.

/13/ K.N. Stevens, "The quantal nature of speech. Evidence from articulatory-acoustic data",Human com.:A unified view,McGrawHill,1972,p.51-66.

UN OUTIL DE PHONETISATION MULTILANGUE

V. AUBERGE[1,2], M. CONTINI[1], D. MARET[1], B. SCHNABEL[1] & H. ZINGLE[3]

1: Institut de la Communication Parlée
Unité associée au CNRS
Institut de Phonétique
Université de Grenoble III
38400 SAINT MARTIN D'HERES
FRANCE

2: Société OROS
ZIRST 38240 MEYLAN
FRANCE

3: Université des Langues et Informatique de Chambéry
73000 CHAMBERY
FRANCE

RESUME

Ce travail résulte d'une collaboration étroite entre informaticiens et linguistes. Il s'agit d'un outil de phonétisation défini dans le cadre de la synthèse multilangue à partir du texte, et conçu pour des applications linguistiques.

INTRODUCTION.

Un certain nombre d'outils de transcription orthographique phonétique du Français ont été développés dans deux principaux buts :

- l'étude de la phonétisation ([9], [8], [3]),
- la synthèse de la parole ([4], [5], [10]).

Dans une optique multilangue, nous avons envisagé la synthèse d'une langue écrite comme un enchaînement séquentiel d'étapes ; à chacune d'elle correspond un module facilement adaptable à la langue considérée : phonétisation, calcul de la prosodie, utilisation d'un dictionnaire (de diphones par exemple).
Nous nous sommes tout d'abord posés le problème du choix de la méthodologie algorithmique qui autoriserait le développement d'un module de transcription commun à toutes les langues orthographiques visées (Français, Allemand, Italien, Espagnol...).
Nous nous sommes particulièrement attachés au développement d'outils conviviaux, permettant un travail de mise au point et d'exploitation dans le cadre d'une équipe pluridisciplinaire : phonéticiens et informaticiens. En effet, il nous a semblé important de pouvoir utiliser les compétences du linguiste en lui proposant un outil qui lui permette de formaliser facilement sa connaissance. En adoptant une telle démarche, nous avons pensé que les règles ainsi obtenues seront à la fois utilisées pour la synthèse, mais aussi pour des études linguistiques relatives à chaque langue.

L'OUTIL DE TRANSCRIPTION

A : Choix méthologique.

Le passage d'une chaîne orthographique vers la chaîne des sons correspondants utilise plusieurs niveaux de connaissances définis chacun par leur unité linguistique minimale (lettre,constituant du mot, mot dans son contexte énonciatif).

Le logiciel élaboré jusqu'à présent s'intéresse aux données linguistiques dont l'unité est la lettre. 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 contextuelles. Il introduit naturellement un ordre local à chaque classe de règles : c'est l'ordre d'application défini par l'ordre d'écriture des règles (on peut représenter ce raisonnement par "exceptions puis règles générales" ou bien "si alors sinon(si...)"). Nous avons donc défini une syntaxe pour concrétiser facilement ce raisonnement : le langage TOPH (voir § B:), et réalisé un ensemble de logiciels d'exploitation de cet outil (figure 1).

Au niveau de la méthodologie algorithmique, nous avons dû choisir entre deux interprétations possibles du mécanisme de transcription :

1.) **Un automate déterministe d'états finis :** dans une étape préalable, il faut expliciter les informations contenues implicitement dans la grammaire. Une règle se développe ainsi en un ensemble équivalent de règles définies exactement sur le vocabulaire d'entrée, indépendantes (le texte d'une règle développée contient les informations nécessaires et suffisantes pour la définir), et en exclusion mutuelle (la grammaire est déterministe). Après avoir construit l'automate d'états finis équivalent au langage constitué de l'ensemble des règles (considérées donc comme des mots de ce langage), on réalise un "pattern matching" entre la chaîne d'entrée et l'ensemble des textes des règles développées.
On obtient donc un algorithme dont la complexité-temps est en O(l), si 'l' est la longueur du texte à transcrire.
Pour construire cet automate, par exemple depuis une grammaire décrite dans le syntaxe TOPH, on la développe :

- selon les **ensembles** ; soit la règle en Français :
("Voyelle")+ s +("Voyelle") = [z]
(c'est-à-dire : 's' entre deux voyelles se transcrit [z])
qui est équivalente à l'ensemble des règles :
(a)+ s +(a) = [z] ; (e)+ s +(e) = [z] ; .... (si "Voyelle"=(a,e,...))
- selon les **opérateurs** ; soit la règle
(a OU e)+ s + = [z] (si 'OU' est l'opérateur logique)
qui est équivalente à l'ensemble des règles :
(a)+ s + = [z] ; (e)+ s + = [z]
- selon l'**ordre d'application** sur les règles ; soit la grammaire décrite par les deux règles:
1: (a)+ s +(a) = [z]
2: + s + = [s]

(c'est-à-dire : si 's' est dans le contexte gauche et droit 'a', il se réécrit [z], sinon il se réécrit [s]),
cette grammaire est équivalente à :
(a)+ s +(a) = [z] ;
("V\a")+ s +("V\a") = [s].
(si "V\a" = "Vocabulaire d'entrée SAUF 'a'")

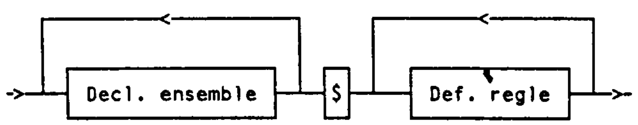
On voit facilement sur ces exemples que la taille de l'automate devient très vite "explosive".

2.) **Un automate transducteur** : la réécriture contextuelle de la chaîne d'entrée est guidée pas à pas par la grammaire. L'ordre induit au transducteur (lecture de gauche à droite de l'entrée), est reproduit naturellement par la partition des règles sur le premier caractère de la sous-chaîne à transcrire.
On obtient alors un algorithme en O(lT), si 'T' représente la taille de la grammaire.

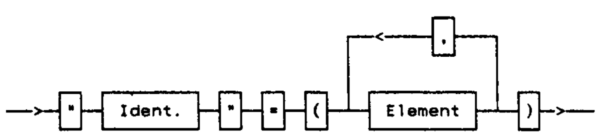
Derrière ces deux démarches, on voit apparaître l'importance prédominante soit des lexiques, soit des règles .
Nous avons finalement choisi la solution transducteur (figure 2), car le gain du facteur 'T' ne nous semble pas rentable devant la complexité de construction de l'automate d'états finis 1.).

B: <u>Le langage TOPH, description syntaxique.</u>
TOPH est un langage LL1.

Grammaire :

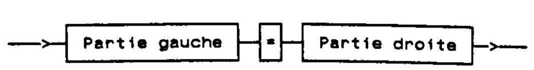


Déclaration d'un ensemble :



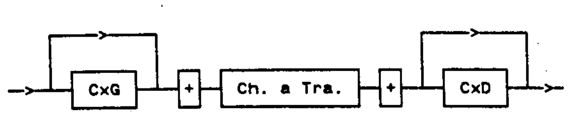
Un identificateur ou un élément d'ensemble sont des chaînes quelconques de caractères.

Définition d'une règle :



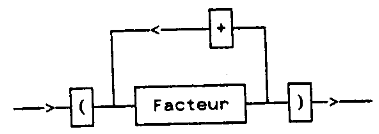
Partie droite :

Partie gauche :

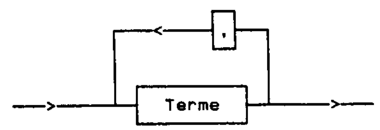


Une chaîne à transcrire (Ve), ou une chaîne transcrire (Vs) sont des chaînes de caractères quelconques.

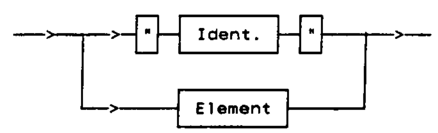
Contexte Droit ou Gauche (même syntaxe) :



Facteur :



Terme :



ETUDE DES GRAMMAIRES

A: <u>La transcription du Français.</u>

A partir d'une grammaire de transcription initiale du Français [9] et des nombreux ouvrages parus([6], [3], [7], [13]...), nous avons affiné les **règles qui se limitent** au mot orthographique.

Les listes d'exceptions ont été extraites du dictionnaire inverse JUILLAND et du dictionnaire inverse du SCRABBLE. Nous avons pu corrigé la grammaire obtenue en l'utilisant pour transcrire le "Frequency dictionary of french words" de JUILLAND, soit près de 5000 formes parmi les plus courantes (une seule forme par base lexicale). Nous allons inclure maintenant l'utilisation de renseignements morpho-syntaxiques, comme par exemple les frontières syntaxiques (liaisons), ou bien la catégorie morphologique. Ainsi, en exemple, l'instanciation à "Verbe" ou "non Verbe" d'une forme finie par -ent décide de sa prononciation en [a˜] ou [ə]. Une analyse morphologique (ou lexicale) peut suffire : exemple 'chantent'/[Sa˜t] et 'souvent'/[suva˜] ; mais il faut parfois remonter à la syntaxe : par exemple pour l'opposition bien connue 'président'/"Verbe"=[presid] ; "non Verbe"=[presida˜].

B: <u>La transcription de l'Allemand.</u>

La conversion d'un texte allemand en chaîne de signes phonétiques correspondante s'effectue en deux pas principaux:
* Le prétraitement morphologique et
* la transcription orthographique - phonétique.

1.) **Le prétraitement morphologique**
opère un découpage des unités lexicales en monèmes et morphèmes (racine, préfixe et suffixe) et la détermination automatique de l'accent.
La procédure d'analyse morphologique est effectuée à l'aide d'un lexique d'environ 150 **préfixes et suffixes**, et un ensemble de 150 **règles pour déterminer la voyelle de la racine qui porte l'accent** (dans tous les cas ou le préfixe ne porte pas l'accent). Le système accomplit ainsi la segmentation morphologique et la détermination de l'accent sur l'unité d'un mot à la fois, par exemple :

Bewunderung ("admiration")

be -	**(préfixe ne portant pas d'accent)**
wunder	**(racine + voyelle portant accent)**
- ung	**(suffixe)**

==>> be-w*under-ung (* = accent)

La méthode choisie [14] se distingue donc aussi bien des systèmes qui se fondent exclusivement sur des lexiques (p.ex. le GRAPHON [8] à Vienne) que des systèmes qui opèrent les analyses morphologiques à 100% par règles
(p.ex. le SYNTEX [12] de Bochum). Cette approche présente trois avantages principaux :
* Les irrégularités de la prononciation de l'allemand dues plutôt à la structure morphologique qu'aux exceptions phonétiques et phonologiques peuvent assez facilement être détectées et définies ; ainsi la fiabilité du système est-elle augmentée.
* Parallèlement le transcripteur reste ouvert à de nouvelles expansions. A chaque instant des nouveaux morphèmes peuvent être introduits dans le lexique.
* En même temps la taille du lexique (qui ne couvre que des préfixes et des suffixes) reste assez raisonnable.

2.) **La transcription**
La chaîne ainsi obtenue, toujours orthographique mais prétraitée, sera transcrite par la suite en signes phonétiques. A l'aide d'un ensemble de 400 règles pour l'allemand la qualité phonétique de chaque lettre est recherchée.
Pour le mot ‹Bewunderung› l'ensemble suivant de règles est appliqué :

b."@."v."*U."n."d."R."U."ng"

("#")+b+ = (b.")
+e+("#") = (@.")

+w+ = (v.")
+u+("CO"+"CO") = (U.")
+n+ = (n.")
("CO")+er+("#") = (R.")

+u+("CO"+"CO") = (U.")
+ng+ = (ng")

Pour les cas qui ne peuvent être transcrits correctement, une liste d'exceptions se trouve dans la grammaire du TOPH. Dans un premier temps nous nous sommes contentés de remplir un **lexique interne** au fur et à mesure que les exceptions se produisent. Par la suite, on isolera ce lexique pour définir un lexique externe.

C: <u>La transcription de l'Italien .</u>

La transcription de l'Italien nécessite un nombre relativement faible de règles en comparaison du Français et de l'Allemand. Ces règles tiennent compte de toutes les distributions possibles de réalisations phonétiques pour lesquelles nous nous sommes référés à la prononciation italienne normative(Toscan cultivé).Ainsi, un -s- intervocalique sera prononcé généralement non voisé, et non pas voisé comme dans tout le Nord de l'Italie. Les phénomènes de Phonétique syntactique n'ont pas été négligés. Ont été pris en considération, en particulier, les phénomènes de sandhi, entraînant notamment une modification de la consonne initiale (allongement), lorsque celle-ci est précédée d'une unité monosyllabique, accentuable ou non :
è bello —>[ɛ bb'ɛllo]

Il en est de même pour les groupements de consonnes aux frontières de mots, pour lesquels des règles différentes ont été adoptées par rapport aux mêmes groupes à l'intérieur d'un mot:
sci —> [S]
la mis cile —> [la mis tS'ile]

CONCLUSIONS

Nous avons pris soin d'écrire un logiciel ouvert, afin de rendre possible toute nouvelle extension, et surtout afin de l'assimiler dans la synthèse en parallèle aux traitements linguistiques nécessaires à la description de la prosodie.

REFERENCES

[1] AHO & CORASICK, "Efficient string machine", A.C.M, June 1975.
[2] V. AUBERGE, "Contribution à la phonétisation des langues alphabétiques", Rapport de D.E.A. - GRENOBLE, 1985.
[3] N.CATACH, "La phonétisation du Français", Edition du C.N.R.S., 1984.
[4] M. DIVAY & M. GUYOMARD, "Contribution et réalisation d'un programme de transcription", Thèse de 3ème cycle - RENNES, 1977.
[5] H.FERVERS, J.LEROUX & L. MICLET, "Programme de transcription orthographique phonémique du Français", Publications E.N.S.T.-D-76003, 1976.
[6] P. FOUCHE, "Traité de prononciation française", Editions Klincksieck, 1969.
[7] V.G. GAK, "L'orthographe du Français", Editions Selac, 1976.
[8] G. JOSEF, G. KAFER & M. KOMMENDA, "Morphologische Analyse im Sprachausgabesystem GRAPHON.", NTG - Fachtagung, München, 1986.
[9] M. LETY, "Transcription orthographique phonétique : un système interpréteur.", Thèse de 3ème cycle - GRENOBLE, 1980.
[10] B. PRATT & G. SYLVA, "PHONTRS.Transcribing french text", Monash University, AUSTRALY, 1967.
[11] B. PROUTS, "Contributions à la synthèse à partir du texte", Thèse 3ème cycle - ORSAY, 1980.
[12] H.W. RUEHL, "A Microprocessor based System for Automatic Conversion of German Text to Speech.", ICASSP 3/3 (1608 - 1611), 1980.
[13] L. WARNANT, "Dictionnaire de la prononciation française", 1962.
[14] H. ZINGLE, "Traitement de la prosodie allemande dans un système de synthèse de la parole.", Thèse d'Etat, Université de Strasbourg II, 1982.

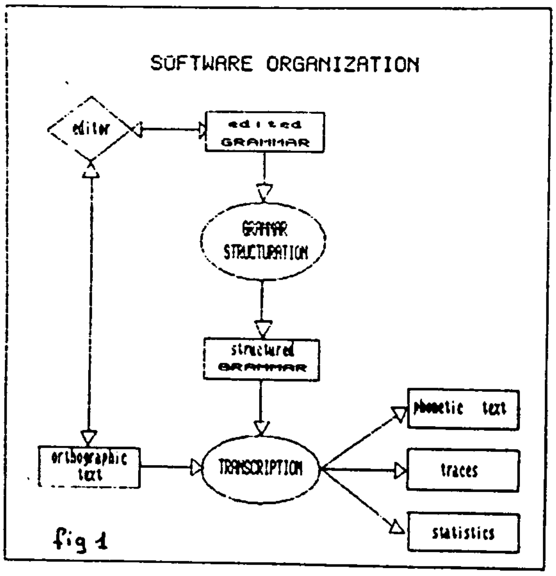
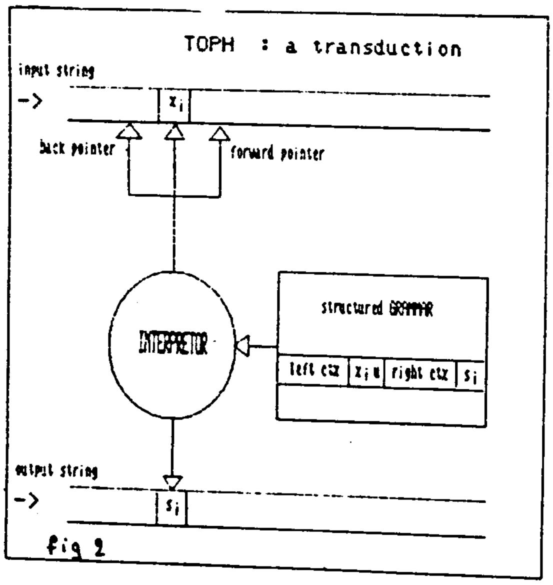

SOFTWARE ORGANIZATION
editor
edited GRAMMAR
GRAMMAR STRUCTURATION
structured GRAMMAR
orthographic text
TRANSCRIPTION
phonetic text
traces
statistics

fig 1


TOPH : a transduction
input string
->
back pointer
forward pointer
INTERPRETOR
structured GRAMMAR
left ctx
right ctx
output string
->
fig 2

FROM SEGMENTAL SYNTHESIS TO ACOUSTIC RULES USING TEMPORAL DECOMPOSITION.

Frédéric BIMBOT, Gunnar AHLBOM, Gérard CHOLLET

ENST Dept. SYC, CNRS UA-820
46 rue Barrault, 75634 PARIS cédex 13, FRANCE.

ABSTRACT

A methodology is proposed to infer automatically acoustic rules that could be used to predict natural spectral transitions for speech synthesis. It adapts ATAL's "temporal decomposition" technique /1/ to compute interpolation functions from phonetically labelled acoustic targets. Coarticulation effects are controlled quite adequately using such a representation. With this methodology, rule-based synthesis will be developped more efficiently for new languages, dialects, speakers with better control of speaking rate, style of speech ...

INTRODUCTION

The automatic generation of "natural" speech from a phonetic transcription is a challenging task. Two main approaches have been proposed: segmental and rule-based. The **segmental** approach (using diphones, demi-syllables, polysons, ...) offers an easy way to intelligible speech. But the segment inventory is speaker dependent and control of timing is a non trivial task.The lack of naturalness could be attributed to uneasy analytic control of speech parameters. A **rule-based** approach is more flexible, gives more insight on the perceptually relevant features of speech, and could be more easily adapted to new speakers. Control of prosody, style of speech, is achieved quite naturally within a unified framework. Unfortunately, this approach requires, so far, a lengthy and art oriented procedure using visual and auditory hand-tuning of the rules.

Our goal is to provide a methodology to move gradually from segmental to rule-based approaches. We propose a number of interactive tools using powerful signal and data analysis techniques to model spectral evolution, infer spectral targets automatically, and generate adequate transitions toward these targets.

SYNTHESIS and COARTICULATION

An acoustic synthesizer is usually controlled by a set of parameters updated at regular time intervals. The parameters are either retrieved from memory (speech restitution and segmental synthesis) or computed from rules. We are concerned here with smooth spectral evolution corresponding to articulatory dynamics. As an working hypothesis, articulatory and therefore spectral targets are assumed. In this paper, coarticulation is referred to as a phenomenon of target undershoot due to contextual effects, speaking rate ...

TEMPORAL DECOMPOSITION

ATAL's technique /1/ decomposes speech into phone-length temporal events which could be interpreted as overlapping and interacting articulatory gestures /2, 3, 4/. Evolution of a sequence of m spectral vectors $[y_i(n)]$ is approximated as a linear combination of m events represented by known functions $Ø_k(n)$ (interpolation functions) with appropriate weights y_{ik}(targets):

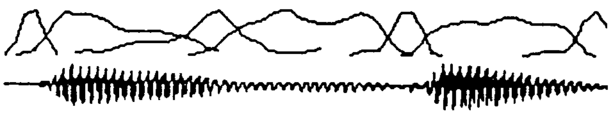
$$y_i(n) = \sum_{k=1}^{m} y_{ik}\, Ø_k(n)$$

The functions $Ø_k(n)$ are constrained to be compact in time: that is zero everywhere except on a segment. The first step of the algorithm consists in finding a good approximation for the localization and the extent of the Ø-functions. Once a set $\{Ø_k\}$ has been found, the corresponding target vectors y_k are computed by:

$$[y_k] = [y_{ik}]\, [Ø_k]^t\, ([Ø_k]\, [Ø_k]^{-1})^t$$

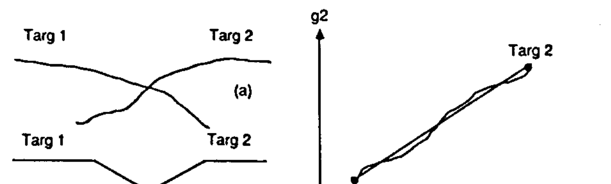
which minimizes the reconstruction error according to a least square criterion.

Iterative refinement can then be performed until no significant improvement is obtained.



Temporal decomposition of the speech segment /ede/.

Ø-functions can be linearly approximated and normalized so that their sum be constant and equal to unity. With this approximation, temporal decomposition of a speech segment correspond to a piece-wise linear trajectory in the parameter space.



Original (a) and linearly approximated (b) interpolation functions
Associated trajectories in the parameter space

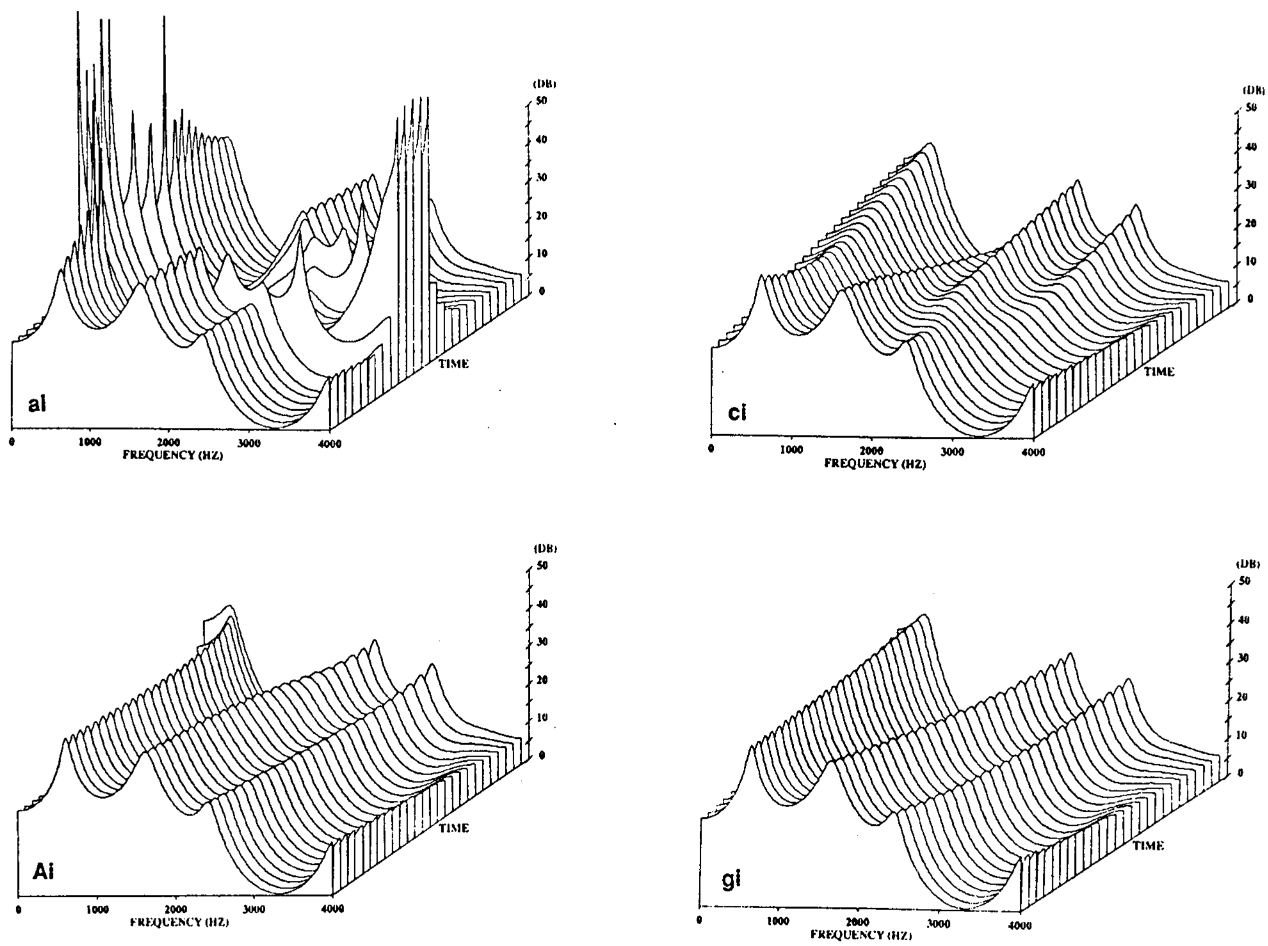


Fig. I Synthetic spectrums associated to a linear trajectory between 2 targets, in different spectral spaces:
ai: auto-regressive coeff, ci: cepstral coeff,
Ai: area parameters, gi: log area ratios.

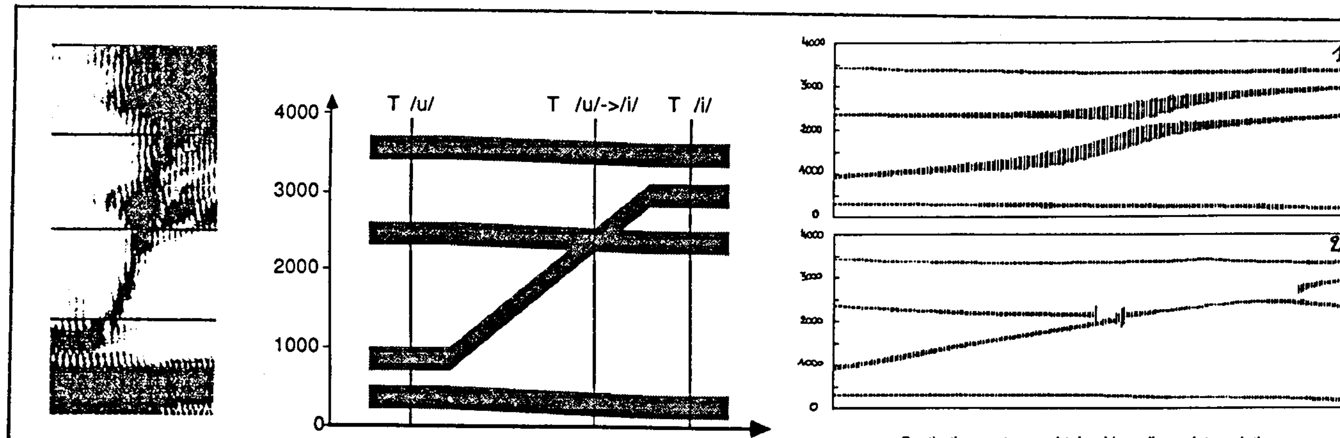
SPECTRAL REPRESENTATIONS

A description of transitions is attempted as a linear combination of spectral parameters. A number of spectral representations could be used for this purpose /5/:

Formant frequencies, amplitudes, and bandwidths (F_i, A_i, BW_i) are often used for speech parameterisation, owing to their physical meaning. However, they necessitate a labelling operation. Moreover, a complex treatment must be performed in order to interpolate spectra with different number of formants. Poles (z_i) and line spectrum pairs (LSP_i) have the same drawbacks.

We therefore investigated the effect of interpolating spectral parameters for several unlabelled spectral representations: LPC autoregressive coefficients (a_i), cepstral coefficients (c_i), area parameters (A_i), reflexion coefficients (k_i), and log area ratios (g_i) /6/.

Auto-regressive coefficients are inadequate as the associated space is not linearly stable. Cepstral coefficients are neither suitable since the mean of two vectors (c_i) gives a spectrum which keeps the peaks of both original spectra. Area parameters seems more convenient, but the interpolated formant trajectories are not quite linear. Reflexion coefficients (k_i) behave adequately with damped resonances. Log area ratios (g_i) are the best parameters we have found so far (see fig. I).



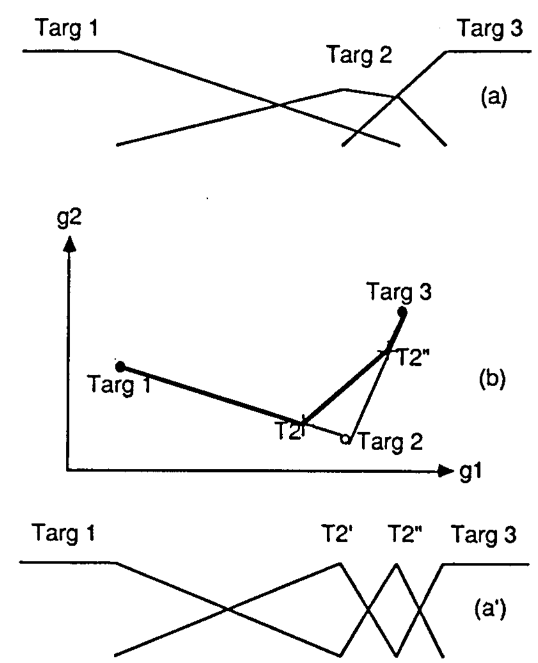
Spectrogram of speech segment /ui/

Typical transition between phonemes /u/ and /i/

Fig. II Synthetic spectrums obtained by a linear interpolation between log area ratios of typical /u/ and /i/ targets. (1) without intervention of any intermediate target, (2) with adjunction of an adequate intermediate target.

COARTICULATION

An analysis of temporal decomposition results reveals the acoustic-phonetic structure of speech. Quasi-stationary segments (fricative, nasal consonant, vowel nuclei) are described with a single function. Transitions are usually described with two overlapping Ø-functions /7/. But some transitions require an extra function $Ø_2$ associated with Targ 2:



Temporal patterns (a) and (a') give different descriptions of the same trajectory (b).
(a) is a best description of the actual articulatory gesture (undershot target)

$Ø_2$ usually describe a highly coarticulated phone with undershoot of the corresponding target. In other cases, the extra function Ø is a correction of the trajectory between targets 1 and 3. This is the case for a rapid front-back movement of the tongue (in such diphones as [ui], [wi], [iu], and [ju]), which correspond to a "crossing formant" configuration /8/. The existence of an intermediate target renders more accurately the spectral transition (see fig. II).

SEGMENTAL SYNTHESIS

Synthesis can be achieved successfully by concatenating stored segments . A set of such segments called "polysons" is chosen in such a way that coarticulation effects accross boundaries are minimized /9/. This is achieved by placing boundaries on spectrally stable sounds (vowels, fricatives, nasal consonants, occlusion of plosives). About 7000 "polysons" were selected for French synthesis. Significant improvement in perceptual quality (intelligibility and naturalness) is achieved with "polysons" synthesis as compared to diphone synthesis. Unfortunately the number of these units is an order of magnitude larger than the number of diphones.

Temporal decomposition can be used to encode "polysons" very efficiently /10/.

RULE-BASED SYNTHESIS

"Polysons" are being classified according to the structure of their Ø-functions /10/. For instance, the temporal patterns of all combinations of a vowel and an unvoiced fricative (/as/, /if/, /us/) are similar.

The archetype of each group can be viewed as a rule to synthesize "polysons" of that group. A "polyson" is therefore reduced to a Ø-pattern type and a set of associated targets.

The edges of "polysons" are quasi-stationary segments, described with a single Ø normalized to unity. The concatenation of "polysons" is restricted to those with matching targets on edges (much like dominos).

The Ø-pattern can be distorted by rules to take care of variations in speaking rate, stress, emphasis... Overlapping and smoothing of Ø-functions at boundaries express the coarticulation effects accross "polysons".

CONCLUSIONS

Temporal decomposition using target spectra can break the complex encoding of these segments. In particular, coarticulation effects are analyticaly explained and modeled. It is demonstrated that these new tools provide an adequate environment in our search for better rules in acoustic speech synthesis.

ACKNOWLEDGMENTS

Some ideas developped in this work where discussed with colleagues from IPO, Eindhoven, during a sabbatical year G. CHOLLET spent there. Contributions of S.M..MARCUS were particularly important in the initial phase.

REFERENCES

/1/ ATAL B.S. **Efficient coding of LPC parameters by temporal decomposition.** Proc. ICASSP-83, 2.6, 81-84, 1983.

/2/ MARCUS S.M., Van LIESHOUT R.A.J.M. **Temporal decomposition of speech.** IPO annual progress report 19, p. 25-31, 1984.

/3/ AHLBOM G., F. BIMBOT, G. CHOLLET. **Modeling spectral speech transitions using temporal decomposition techniques.** ICASSP, Dallas, 1987.

/4/ CHOLLET G., GRENIER Y., MARCUS S.M. **Temporal decomposition and non-stationary modeling of speech.** EUSIPCO, The Hague, 1986.

/5/ SCHAFER R.W., RABINER L.R. **Parametric representations of speech.** From: Speech Recognition, REDDY R. (ed.), 1975.

/6/ VISWANATHAN R., MAKHOUL J. **Quantization properties of transmission parameters in linear predictive systems.** IEEE Trans. ASSP 23, pp. 309-321, 1975.

/7/ OHMAN S.E.G. **Coarticulation in VCV utterances: Spectrographic measurements.** JASA 39, pp. 151-168, 1966.

/8/ CHAFCOULOFF M., CHOLLET G., DURAND P., GUIZOL J., RODET X. **Observation and modeling of the variability of formant transitions using ISASS.** IEEE-ICASSP, Denver, 1980.

/9/ CHOLLET G., GALLIANO J. F., LEFEVRE J. P., VIARA E. **On the generation and use of a segment dictionnary for speech coding, synthesis and recognition.** IEEE-ICASSP, Boston, 1983.

/10/ BIMBOT F., CHOLLET G., MARCUS S. M. **Localisation et representation temporelle d'evenements phonetiques: applications en etiquetage, segmentation et synthese.** JEP-86, Aix-en provence, 1986.

A New Program for Manipulation of Natural Speech :
---- Interpolation Between Two Natural Utterances ----

Maximilian Hadersbeck

Institut für Phonetik und sprachliche Kommunikation
der Universität München
Schellingstr. 3, 8000 München 40, F.R.G.

ABSTRACT

For phonetic experiments it is very important to be able to manipulate distinctive features in natural utterances by not loosing the natural sound of the utterance.
In this paper the description of a computer program is delivered which allows interpolation between two natural utterances by means of Spectral Envelope Interpolation. The program produces high quality synthetic utterances, where speech parameters like pitch, intensity and formant structure of one speech utterance can be adopted to the same parameters of another utterance.
You can produce a natural sounding utterance continuum starting with the unmanipulated initial utterance towards a final utterance with the desired degree of manipulation. With this kind of manipulation you can for example change speaker identity, sentence intonation and stress of a natural utterance.

INTRODUCTION

This program is a new solution to interpolation between two natural utterances. The program produces higher quality speech utterances than a former program, developed by Simon [1], [2] at this institute.
Simon describes the program as following :

"Contours of speech parameters such as pitch, intensity or formant structure can partially or totally be imposed on utterances, thus changing speaker identity, sentence intonation, stress or other psychophysical parameters. All manipulation can be done in discrete steps from the unmanipulated initial utterance to the final utterance with the desired degree of manipulation." (Simon (1984))

The program which I developed allows speech utterances to be manipulated in the same way as Simon described, but I introduced a new interpolation method in this program. The interpolation method is a means of "Spectral Envelope Interpolation" :
The spectra of the two utterances are calculated pitch-synchronously and according to the mode of interpolation and number of discrete interpolation steps, out of the two spectra a new spectrum is geometrically developed. From this spectrum the time signal of the new speech utterance is produced by a means of Inverse Discrete Fourier Transformation.
A continuum between two utterances can be developed by interpolating between the following parameters :

Spectral and Intonation Interpolation

The spectral shape and the intonation of speaker one's utterance is interpolated towards the spectral shape and intonation of speaker two's utterance.

Spectral Interpolation (the intonation of utterance one remains unchanged)

It is interpolated between the spectral shape of the two utterances. All the utterances of the continuum have the intonation of utterance one.

Intonation Interpolation (the spectrum of utterance one remains unchanged)

It is interpolated between the intonation of the two utterances. The spectral shape of the utterances is not changed.

THE PROGRAM [1] :

Logarithmic Interpolation :

The user can choose between a linear and a logarithmic interpolation method.
The logarithmic interpolation method has the advantage that the steps from the intonation of the first utterance towards the second are not simply analytically defined, but they are fitted to the properties of listener's speech perception

Spectral Envelope Interpolation :

The interpolation method between every single pitch period of the two utterances is a "Spectral Envelope Interpolation" method. This method avoids some disadvantages of the interpolation method, introduced by Simon [1].
One disadvantage was distortion in the synthetic speech signal if the two utterances are very different in their F0 structure.

Input and Output :

The inputs of the program are the two speech utterances, with a description of the unvoiced/ voiced/ pause structure. The next input is the number of discrete interpolation steps between the two utterances and the mode of interpolation. The last input is the kind of interpolation : Linear interpolation or logarithmic interpolation.

The outputs of the program are a continuum of natural sounding utterances with their unvoiced/ voiced/ pause description in accordance to the input.

Restrictions on the Phonetic Structure of the Two Utterances :

The two utterances must have the same phonetic unvoiced/ voiced/ pause structure.
An example : Given is the utterance "MAX". Its unvoiced/ voiced/ pause structure is :

```
 P   V   V   P   U
/    M   A      KS /
```

with P for pause, V for voiced, U for unvoiced.

The second utterance must have the same unvoiced/ voiced/ pause structure:

```
 P   V   V   P   U
```

for expample : "MIX"

The program interpolates now between the first segment (Pause in the example) of utterance one and the first segment in utterance two, the second segments (voiced segment in the example), the third and so on.
The program cannot interpolate between utterances with different phonetic structure. The only parameter which can vary between the utterances is the information within the coinciding segments :
The segments can have totally different length, F0 Curve, energy distribution, they can be spoken for example from different speakers. Between these parameters the program can interpolate.

The Interpolation Between the Two utterances Within the Speech Segments.

Pause- Segments :

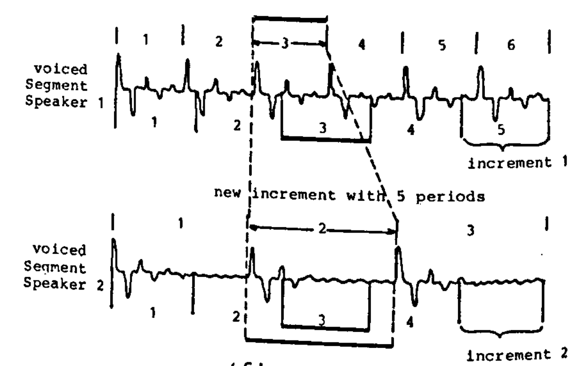
To interpolate between Pause segments, the speech datas are transformed with small modifications. The duration of the Pause segments are adopted to the interpolation step.

Unvoiced- Segments :

According to the interpolation step, the duration of the unvoiced segment is adopted. The speech data for the new utterance are calculated from the time signal of the two utterances, without any spectral transformation.

Voiced- Segments :
Determination of The Actual Pitch Period :

In the first step the program counts the number of pitch periods within the voiced segment for both utterances. We call the number of pitch periods in utterance one IANZ1, those in utterance two IANZ2. According to the interpolation step, the number of pitch periods for the new utterance (IANZX) is calculated. (see Formula 2 in Appendix)
Now two increments are developed to move through the voiced segment : INCR1 (utterance one) and INCR2 (utterance 2). With that increment you can step through the voiced segment of utterance one and utterance two, meeting IANZX pitch periods in the segments of utterance one and in utterance two.



(figure 1)

An example: (look at figure 1)

Utterance one has 6 pitch periods and utterance two has 3 pitch periods.
The interpolation step demands 5 pitch periods for the new utterance.
Now you calculate an increment for speaker one and two, to step through the utterances to meet 5 pitch periods in utterance one and utterance two.
With that increment you step through the voiced segments, and the current increment pointer determines the pitch periods in utterance one and utterance two, between which the "Spectral Envelope Interpolation" is performed.
In the example an interpolation is performed between pitch period number 3 of utterance one and pitch period number 2 of utterance two, to calculate the new pitch period number 3.

The "Spectral Envelope Interpolation" :

The aim of this method is, to calculate out of two pitch periods, with their F0 value, a new pitch period which has an F0 value in accordance to the interpolation step.
I decided to take the Discrete Fourier Transformation and Discrete Inverse Fourier Transformation to reach this aim [2].

The Algorithm :

Step 1 :
Calculate the N spectral lines U1p(n) of the time signal u1p(n), n=1,...,N in utterance one within pitch period p

Calculate the M spectral lines U2q(n) of the time signal u2q(n), n=1,...,M of utterance two within pitch period q (See Formula 1 in the Appendix).

Step 2 :
Calculate the number Q of neccesary spectral lines in the new period according to the interpolationstep.
(See Formula 2 in the Appendix)

Step 3 :
Calculate the Q spectral lines W1p(n) for utterance one out of the N spectral lines U1p(n) via the "Spectral Envelope Interpolation"

Calculate the Q spectral lines W2q(n) for utterance two out of the M spectral lines U2q(n) via the "Spectral Envelope Interpolation".
(See Subroutine ENVINT in Appendix)

Step 4 :
Calculate the new spectral Lines Xf(n), n=1,...,Q via an interpolation between W1p(n) and W2q(n).
(see formula 3 in the Appendix)

Step 5 :
Calculate the time signal for period f xf(n), n=1,...,Q for the new utterance out of the Q spectral lines Xf(n)
(see formula 4 in the Appendix)

IMPLEMENTATION :

The program is implemented in the High Level Language FORTRAN 77 on a Digital Equipment Corporation Computer PDP11/73 [3].
The main emphasize in the program lies on a very good readable form. Its subroutines are clearly defined and a programmer can understand very easy the work of the routines.
Because of this, the computation time is rather long : On a PDP 11/73 it takes for example 15 minutes to interpolate between two utterances of 5 seconds length.

SUMMARY :

I have descibed in this paper a computer program, which is able to produce a continuum of natural sounding utterances. The utterance continuum starts with an unmanipulated utterance towards a manipulated utterance. The kind of manipulation can be chosen manually via program input.

ACKNOWLEDGEMENTS :

The work was sponsored by DIGITAL EQUIPMENT CORPORATION GmbH Munich, FRG with a complete PDP11/73 computer system, including Software and Analog/Digital periphery [3] [4].

REFERENCES :

[1] Simon, Th. : Manipulation of natural speechsignals according to the speech parameters of different speakers, Forschungsberichte des Instituts fuer Phonetik und Sprachliche Kommunikation der Universitaet Muenchen (FIPKM) 17 (1983) page 233-245
[2] Rabiner L. Gold B. :Theory and Application of Digital Signal Processing, Prentice Hall (1975)
[3] Hadersbeck M. :Sprache und Klang Eine MICRO 11/73 als Speechprocessingmachine, 9. DECUS Muenchen e. V. Symposium, Suttgart (1986)
[4] Hadersbeck M. :Digitale Sprachverarbeitung unter Micro RSX, 10. DECUS Muenchen e. V. Symposium, Berlin (1987)

APPENDIX :

Formula 1 [2] :

$$(2)\quad Up(n) = \sum_{l=0}^{N-1} up(l)\cdot e^{-i(2\pi/N)ln}$$

for n=1,...,N

Discrete Fourier Transformation

Formula 2 :

$$Q = N + FF * (M - N)$$

where :
FF ... factor of interpolation
(0.0 <= FF <= 1.0)

Formula 3 :

$$Xf(n) = W1p(n) + FF * (W2q(n) - W1p(n))$$
for n=1,...,Q

where :
FF ... factor of interpolation
(0.0 <= FF <= 1.0)

Formula 4 [2] :

$$(3)\quad xf(n) = 1/N \sum_{l=0}^{N-1} Xf(l)\cdot e^{i(2\pi/N)ln}$$

for n=1,...,Q

Discrete Inverse Fourier Transformation

```
      Subroutine ENVINT
                (SPECIN,ANZIN,SPECOU,ANZOU

C "SPECTRAL ENVELOPE INTERPOLATION METHOD
C  Input :
C  SPECIN(I), I=1,,,ANZIN
C       ....  Spectral Lines Input
C  Output :
C  SPECOU(I), I=1,,,ANZOU
C       ....  Spectral Lines Output

   DIMENSION SPECIN(1),SPECOU(1)
   INTEGER*2 ANZIN,ANZOU

   XIN=2*3.1459265/ANZIN   ! angle input
   XOU=2*3.1459265/ANZOU   ! angle output
   QQ2=ANZOU/2    !ANZOU/2 spectral lines
   DIN=0.0        !Increment counter IN
   DOU=XOU        !Increment counter OUT
   IND=1          !spectral line counter
   SPECOU(1)=SPECIN(1)

   DO 1 I=1,QQ2
2  IF (DOU .GE.  DIN
      .AND.  DOU .LT.  DIN+XIN) GOTO 3
   IND=IND+1
   DIN=DIN+XIN
   GOTO 2
3  GRAD=(DOU/DIN)/XIN
   SPECOU(I+1)=
   (SPECIN(IND+1) - SPECIN(IND))*GRAD +
         SPECIN(IND)
   DOU=DOU+XOU
1  CONTINUE

   RETURN
   END
```

SYNTHESE DE LA PAROLE PAR POINTS-CLES : PREMIERS RESULTATS

AGNES MANTOY

Laboratoire "Image et Parole" et Laboratoire de Phonétique (D.R.L.)
Université Paris 7 - Paris

RESUME

Compte tenu de la redondance inhérente au signal de parole et de la pertinence de certains événements, il est possible de reconstituer un signal de qualité acceptable à partir d'un jeu de paramètres attachés à certains "points-clés" du signal d'origine. Ce principe est mis en oeuvre ici sur des phrases simples de français standard : les points-clés sont recherchés sur les représentations temporelle et spectrale du signal. Entre ces points, les coefficients de réflexion nécessaires à la synthèse par prédiction linéaire sont ensuite calculés par interpolation.

INTRODUCTION

La parole est dotée d'une redondance importante à quelque niveau que se situe l'analyse et en particulier au niveau acoustico-phonétique : la présence d'un phone à un instant donné influe sur la réalisation acoustique des phonèmes environnants.

Ainsi le signal de parole est constitué de segments stables ou quasi-stables et d'autres, transitoires, reflétant un changement plus ou moins important et plus ou moins rapide de la source sonore et/ou des articulateurs. On peut y repérer, quelquefois non sans difficultés, des discontinuités majeures comme début et fin de voisement, début et fin vocalique, début et fin de friction etc ... (cf Abry & al. [1]).

Les synthétiseurs existants exploitent la redondance du signal pour réduire le débit d'information très élevé du signal d'origine en maintenant son intelligibilité avec une bonne qualité, mais il est encore possible de réduire ce débit d'information en utilisant les propriétés acoustico-phonétiques du signal, c'est-à-dire en tenant compte de ces événements entre lesquels le signal évolue

Certains auteurs ont déjà travaillé dans ce sens, en particulier Olive & Spickenagel [4], et l'on se propose de reprendre ce travail sur des courtes phrases de français standard, de façon plus systématique et en partant de considérations plus phonétiques que techniques.

CORPUS

Il est constitué d'une quarantaine de mots de type CVCV insérés dans la phrase porteuse "C'est ---- ça". La deuxième voyelle est toujours /a/, la première étant /i/, /a/, /u/ ou /ã/. Les consonnes employées sont les suivantes : /t/, /k/, /b/, /d/, /n/, /s/, /ʃ/, /v/, /z/, /l/ ou le glide /j/. Ce corpus a été enregistré en chambre sourde et dans un ordre aléatoire par un locuteur masculin.

Les phrases ont été ensuite numérisées à 16 kHz sur 12 bits. L'analyse LPC, effectuée par tranches de 16 ms, fournit un jeu de 14 coefficients de réflexion auxquels il faut ajouter le gain et le pitch. Après lissage du pitch et du gain, on procède à une synthèse LPC qui restitue un signal "de base" reconstruit directement à partir des coefficients de réflexion d'origine.

METHODE

Le traitement consiste tout d'abord à rechercher sur les représentations temporelle et spectrale du signal de base des "points-clés", c'est-à-dire des points indispensables à la reconstitution d'un signal de qualité. A ces points-clés sont associés les coefficients de réflexion de la tranche d'analyse correspondante. Entre ces tranches, les paramètres sont calculés par interpolation linéaire de l'arc sinus des coefficients de réflexion. La synthèse LPC effectuée sur ces jeux de paramètres fournit alors de nouveaux signaux "interpolés".

La qualité de la synthèse dépend bien sûr du nombre de points-clés retenus, de leur emplacement mais aussi du type d'interpolation effectuée tant sur les coefficients de réflexion que sur les paramètres prosodiques. Dans un premier temps cependant, nous avons choisi de limiter notre étude à l'évolution des paramètres relatifs au conduit vocal indépendamment de ceux associés à la source sonore : nous avons donc conservé, pour chaque signal "interpolé", le gain et le pitch du signal de base. Par ailleurs, les travaux de Nordstrand & Öhman [3] ont montré que l'interpolation linéaire en arc sinus des coefficients de réflexion donne de meilleurs résultats que d'autres méthodes (interpolation linéaire des fonctions d'aire ou des coefficients "LAR" (Log Area Ratio), par exemple). Il va sans dire que lorsque le problème de la recherche des points-clés sans prise en compte de l'évolution des paramètres prosodiques sera résolu, ceux-ci devront être réintroduits et l' on sera amené à revoir le type d'interpolation à effectuer.

SEGMENTATION

Elle est effectuée non pas sur le signal original mais sur le signal de base reconstruit après analyse-synthèse LPC puisque, du fait de l'analyse sur des tranches de 16 ms, il peut y avoir un décalage entre la localisation d'un événement acoustique donné sur l'original et sur le signal synthétisé. Par ailleurs, on utilise le spectrogramme et l'édition du pitch et de l'intensité comme aide à la segmentation.

La segmentation consiste ici à marquer les frontières des zones stables, aucune décision n'étant prise pour les segments transitoires quant à leur appartenance à l'un ou l'autre des phones qui les entourent. Les phones discontinus tels que les occlusives sont subdivisés en deux segments : le premier (silence ou voisement) correspondant à la tenue de la consonne, le second, bruité, à son relâchement.

RECHERCHE DES POINTS-CLES

Si l'on suppose que deux points-clés par phone en moyenne sont nécessaires à la reconstitution du signal (cf Heller [2]), alors plusieurs stratégies sont possibles parmi lesquelles les deux suivantes :

(i) définir les points-clés comme les milieux des parties stables et des transitions : ce marquage indique bien les cibles à atteindre mais ne rend pas compte de la durée respective de ces parties stables et de ces parties transitoires. De plus, la décision concernant la localisation du point situé dans la transition n'est pas toujours facile à prendre.

(ii) définir les points-clés comme les extrémités des zones stables : les cibles à atteindre ainsi que la durée pendant laquelle elles sont tenues sont bien prises en compte. En outre, ces points, entre lesquels les coefficients de réflexion sont interpolés, présentent l'avantage d'être relativement sûrs. Ceci suppose que les formants des zones de transition sont des courbes continues et monotones d'une cible à l'autre, hypothèse qui devra être affinée.

C'est cette deuxième méthode qui a été retenue. Cependant, le résultat, à l'audition du signal et sur sa représentation graphique, n'est pas toujours satisfaisant. En effet, les courbes des formants dans les parties transitoires n'ont pas une pente constante. De plus, il semble que ces transitions doivent être interprétées plus comme des ajustements des articulateurs en jeu, avec les erreurs que cela comporte, que comme un déplacement monotone de ceux-ci d'une cible à l'autre ("overshoot"). Si tel est le cas, il faut donc déterminer les variations pertinentes dans le mouvement des articulateurs, c'est-à-dire pour nous, les variations pertinentes des coefficients de réflexion dans les zones de transition.

Pour tenter de résoudre ce problème, nous avons également sélectionné un point-clé à l'intérieur de la transition. Le nombre moyen de points-clés par phone est alors de trois et non plus de deux, ce qui rend la synthèse de bien meilleure qualité mais aussi plus coûteuse en stockage de données. Cette méthode conduit à

(i) se limiter, sur les voyelles, uniquement au segment où la structure formantique est quasiment constante,

(ii) marquer dans la transition le début (dans le cas C-V) ou la fin (dans le cas V-C) de la structure vocalique (établissement ou relâchement de la voyelle).

Sont soumis au même traitement que les voyelles tous les sons vocaliques (présentant une structure formantique) tels que les consonnes nasales ou latérales et les glides ainsi que la consonne approximante /v/.

En fait,le nombre réel de points-clés dépend de la composition de la séquence sonore : une portion de signal ne comportant que des sons vocaliques par exemple, a toujours une structure formantique, variant dans le temps, mais ininterrompue. Il n'y a donc à retenir pour chaque phone que les deux points-clés situés aux extrémités de la partie stable. Certains de ces sons vocaliques ont une partie stable très brève voire inexistante (le glide /j/ ou la latérale /l/ par exemple) et un seul point suffit dans ce cas.

COEFFICIENTS DE REFLEXION ET POINTS-CLES

L'originalité de notre méthode consiste à nous appuyer aussi sur l'évolution temporelle des deux premiers coefficients de réflexion. Le rapport entre représentation temporelle et spectrale du signal et représentation temporelle des coefficients de réflexion nécessite une étude approfondie qui sera menée ultérieurement. Cette relation n'est pas évidente, par exemple, dans les syllabes non accentuées et entièrement voisées qui sont difficiles à segmenter et dont les coefficients de réflexion présentent des fluctuations assez déconcertantes. Toutefois, on constate que la visualisation des coefficients permet le plus souvent, en cas de doute sur deux tranches adjacentes, de sélectionner l'une d'entre elles comme point-clé.

Par ailleurs, du fait du principe même de l'analyse LPC, notre méthode se heurte à un problème de résolution temporelle : deux événements successifs choisis sur la représentation temporelle du signal peuvent éventuellement se situer dans la même tranche d'analyse ou dans deux tranches adjacentes : Cela se produit à la frontière d'une consonne occlusive suivie d'une voyelle, où nous sommes toujours amenés à retenir des point-clés dans plusieurs tranches contiguës, l'un marquant la fin de la tenue de l'occlusive, l'autre son relâchement, le troisième la fin du relâchement de la consonne et le début de l'établissement de la voyelle. Il est possible aussi que deux événements distincts soient situés dans la même tranche d'analyse et contribuent alors, ensemble, aux valeurs que prennent les coefficients. On devra, dans une étape ultérieure (segmentation automatique par exemple), tenir compte de cette difficulté.

APPLICATIONS

Les figures 1 et 2 représentent l'oscillogramme et le spectrogramme des phrases "c'est douta ça" et "c'est vida ça" Elles permettent de comparer les signaux de base synthétisés à partir des coefficients de réflexion originaux (66 jeux de paramètres) avec ceux reconstruits à partir des coefficients interpolés.

Ces signaux "interpolés" ont été resynthétisés à partir respectivement de 23 et 19 points-clés Le nombre de points retenus reste donc relativement important, mais en contre-partie la qualité de ces signaux est très bonne et il est difficile de les distinguer à l'oreille des signaux de base.

CONCLUSION

Bien que le type de synthèse étudié dans cette communication résulte en une compression du débit d'information, elle apparaît surtout comme un outil bien adapté à la mise en évidence de l'importance perceptuelle d'événements acoustiques dans la compréhension de la parole. On peut penser que les événements qui doivent être retenus sont le reflet de changements articulatoires essentiels lors de la production. Leur détermination nécessite une étude systématique des cas où apparaissent les phénomènes de coarticulation, étude qui devra être menée sur la production de plusieurs locuteurs. Une analyse plus fine du comportement des coefficients de réflexion sera alors possible et permettra sans doute de distinguer les variations pertinentes de celles qui ne le sont pas et de les rapprocher des changements intervenus dans les représentations tant temporelle que spectrale du signal.

REMERCIEMENTS

Tous les traitements numériques ont été effectués sur système GenRad en TSL (Time Series Language). Nous remercions la société GenRad France et en particulier Messieurs Zeguer et Stohl pour le soutien qu'ils nous ont apporté.

REFERENCES

[1] Abry C. & al. - propositions pour la segmentation et l'étiquetage d'une base de données des sons du français - 14 ème JEP GALF 1985, p. 156-163.

[2] Heller J. - Optimized frame selection for variable rate synthesis - IEEE ICASSP 1982, p. 586-588.

[3] Nordstrand L. & Öhman S.E.G. - Computer resynthesis of speech on phonetic principles - Lund Univ. W.P. n° 19, 1980, p. 74-79.

[4] Olive J.P. & Spickenagel N. - Speech resynthesis from phoneme-related parameters - JASA vol. 59, n° 4, 1976, p. 993-996.

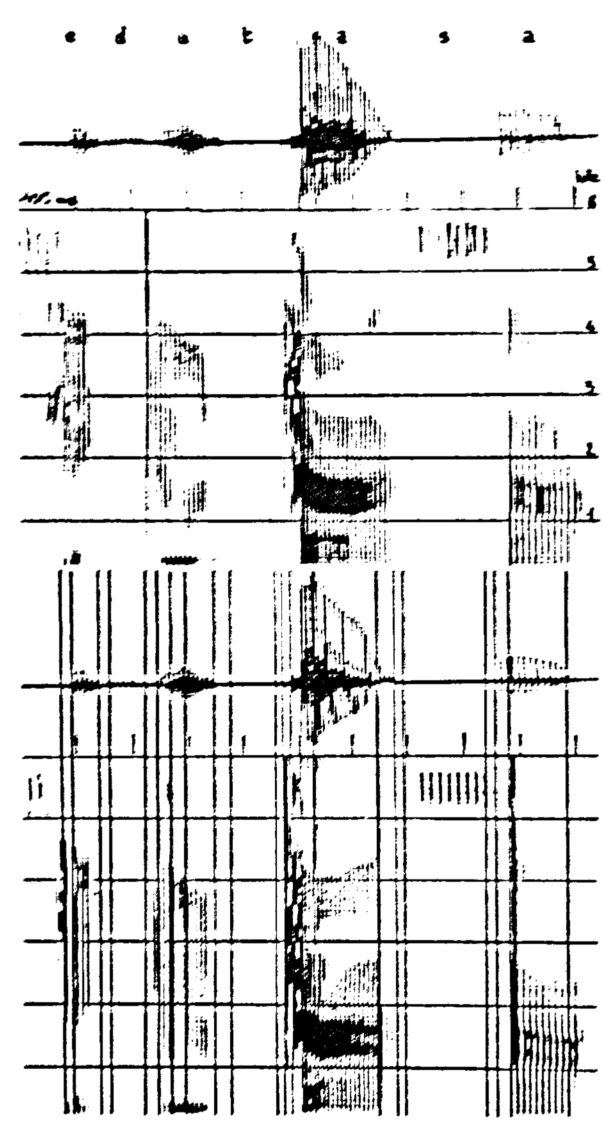


Fig 1 Oscillogramme et spectrogramme de /sedutasa/. En haut : signal de base, en bas : signal "interpolé".

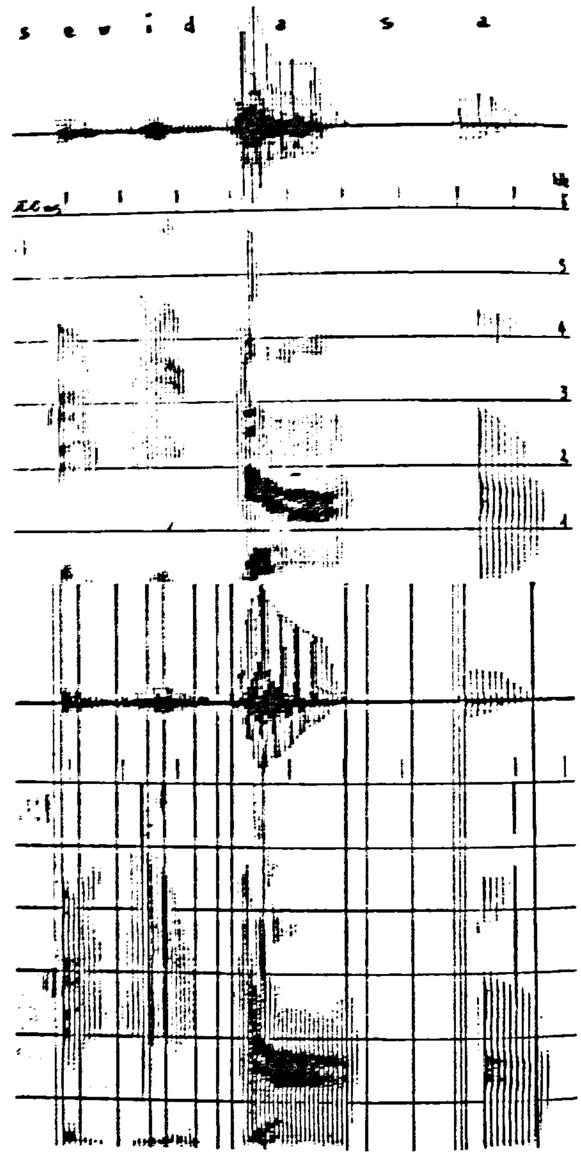


Fig 2 Oscillogramme et spectrogramme de /sevidasa/. En haut : signal de base, en bas : signal "interpolé"

EFFECTS OF CONTEXT AND LEXICAL REDUNDANCY ON CONTINUOUS WORD RECOGNITION

PETER J. SCHARPFF

Dept. of Linguistics/Phonetics Laboratory,
Leyden University, P.O. Box 9515,
2300 RA Leiden, The Netherlands

ABSTRACT

Word recognition research typically focusses on the recognition of isolated words. Yet in actual speech perception the correct or incorrect recognition of earlier words will be crucial to the recognition of later words in the sentence and vice versa. Using an ongoing gating technique, the effects of lexical redundancy (monosyllabic vs. polysyllabic words) and speech quality (synthetic speech, degraded natural speech, high quality natural speech) on word recognition were investigated.
The results reveal that sentences composed of short words are more difficult to understand than sentences with longer words, as can be predicted by e.g. the Cohort model of word recognition. Also, it appeared that when a word could not be recognized instantaneously (as often occurs in low quality speech), chances of a postponed recognition on the basis of following context abruptly decrease when more than 4 words (or 7 syllables) have elapsed. Such delayed recognition of earlier words typically occurs at constituent boundaries.

INTRODUCTION

When a listener hears a sequence of sounds like "Inabankmanagersoff..." he can't be sure yet whether this would be the beginning of the sentence

(1) In a bankmanager's office law and order must rule.

or

(2) In a bank, managers offer a lot of service to customers.

A decision as to how the incoming sounds should be divided into words can be made only when we have heard enough of the following context to solve the ambiguity. Such ambiguities pose problems to the listener, especially when the segmental quality of speech is poor, e.g. as a result of background noise or due to the fact that speech is produced by a machine.
The number of alternative interpretations that the listener must keep in mind during the process of recognition can be very large, and the listener will need relatively much of the following context to solve an ambiguity. These kinds of problems are caused by the fact that the listener does not know where to place word boundaries. When giving away those boundaries we will help the listener to solve ambiguities and to integrate the sounds he has already heard. This can be done by means of prosodic word boundary markers like a pitch rise at the end of a phrase, a non-final pitch fall between two rises or a speech pause (all three accompanied by lengthening of the preceding syllable).
In previous research (see [1] and [2]) it was shown that it is possible to reduce the negative effects of poor segmental intelligibility by placing a clear speech pause after, for instance, every related group of words. In this research the recognition percentage increased with 10 points as a result of pauses edited into the speech.
When prosodic boundary markers are to be edited in continuous speech, these have to be inserted at those places that help the listener recognize the speech as much as possible.
Not only does reduced speech quality affect the intelligibility but also word length can play an important role in the delay of word recognition. Long (polysyllabic) words will be recognized early relative to their word length as opposed to short (monosyllabic) words. This effect can be explained as a result of the inherent lexical redundancy of longer words. Such redundancy is generally absent in short words. When a listener hears the sound sequence "eleph..." he will undoubtedly recognize (under perfect listening conditions) the word "elephant" even if he has not heard the final syllable yet, because there is no other (monomorphematic) word in his vocabulary that begins with this sound sequence. The moment that a listener has heard enough of the sound material to determine which word it will be, is called the recognition point of that word. It will be clear that shorter words contain far less or even no lexically redundant material. The lack of redundancy in words results in a shift of the recognition point towards, or even beyond the word end. This tendency will even be increased by the effect of degraded speech quality. In such cases a listener will need more of the following context to solve his recognition problems.
In an experiment systematically varying word length and speech quality we have examined the following questions:

a. To what extent does word length (or lexical redundancy) influence the recognition of words in connected speech?
b. What is the maximal stretch of following context that a listener may use to facilitate the recognition of a word?

METHOD

When we want to establish the positions in a sentence where most of the recognition problems arise and how long such problems may persist for a listener, we must be able to trace responses from the listener from moment to moment. This is possible when we use a gating technique in presenting stimuli to subjects. The technique used in this experiment presents fragments of sentences to subjects that are lengthened on each following presentation, until eventually the listener has heard the whole sentence. The length of one increment used in this particular experiment is a speech fragment that begins in the middle of the vowel of a lexically stressed syllable and ends in the middle of the vowel of the next stressed syllable (roughly comparable to a 'foot'). The first fragment is of course from the sentence onset to the middle of the vowel from the first stressed syllable.
For each sentence three versions were constructed with different speech qualities: hi-fi natural speech, natural speech degraded by amplitude-modulated white noise, and diphone synthesis using a Philips MEA 8000 speech chip. The rationale behind including degraded natural speech was that we wished to check whether the same type of errors were obtained under poor speech quality irrespective of the precise type of degradation.

MATERIAL

Pairs of sentences were constructed in which we varied poly- and monosyllabic words in the same syntactical structure and with a similar meaning. For example:

(3) Een knecht vond het kind op de stoep van zijn huis.
(A servant found the child on the doorstep of his house.)

and

(4) Een agrarier ontdekte de vondeling in een weiland nabij zijn boerderij.
(An agrarian discovered the foundling in a field near his farm.)

Thirty subjects were asked to listen to the stimuli each time guessing what word the word fragment they heard last would be the beginning of. They had to type their responses into a computer, that was programmed to analyse the answers on what was correct and what was not. After having been informed what words had been correct, the subjects listened again to the sentence now lengthened with one 'foot' of context, corrected their earlier response when necessary and added what they had recognized of the newly heard sound sequence. All responses of the subjects throughout all stages of the experiment were stored in computer memory.

RESULTS AND CONCLUSIONS

Because in the material only content words were systematically varied with respect to word length, we analysed only the responses to those words.
Turning to the first question of the experiment, whether word recognition is more difficult in the versions with short words than in the versions with long words, we find that the longer words were indeed recognized better than the short words: 96% versus 92.5% correct. The difference is fairly small. However when we look at table I, we see that the difference in word recognition of long and short words is substantially larger for the synthetic speech quality:

	short words	long words
hifi	99.9%	99.8%
noise	95.5%	97.7%
synthetic	82.0%	90.4%
mean	92.5%	96.0%

Table I. Percentage correct recognized content words after final presentation. N [short words] = 2400; N [long words] = 2400.

There is no difference at all between the word recognition of long and short words under hifi speech quality. The versions with noise were still recognized better than the synthesized versions, because, as we analysed, we found that listeners get used to the noise; learning effects were much smaller for synthetic speech. In pilots the noise level masking the human speech was adjusted so as to make degraded human speech as (un)intelligible as the diphone synthesis. However, due to the much shorter exposure times in the pilots, no differences in learning effects were discovered before the main experiment.
The differences between the three speech qualities were all significant. This leads us to conclude that words are more difficult to recognize when speech quality gets worse. Moreover, it appears that recognition of short words suffers more from the negative effect of degraded speech quality than that of long words.

The next question to be answered concerns the maximal stretch of following context that a listener may use to facilitate the recognition of a word. Consider the next figure:

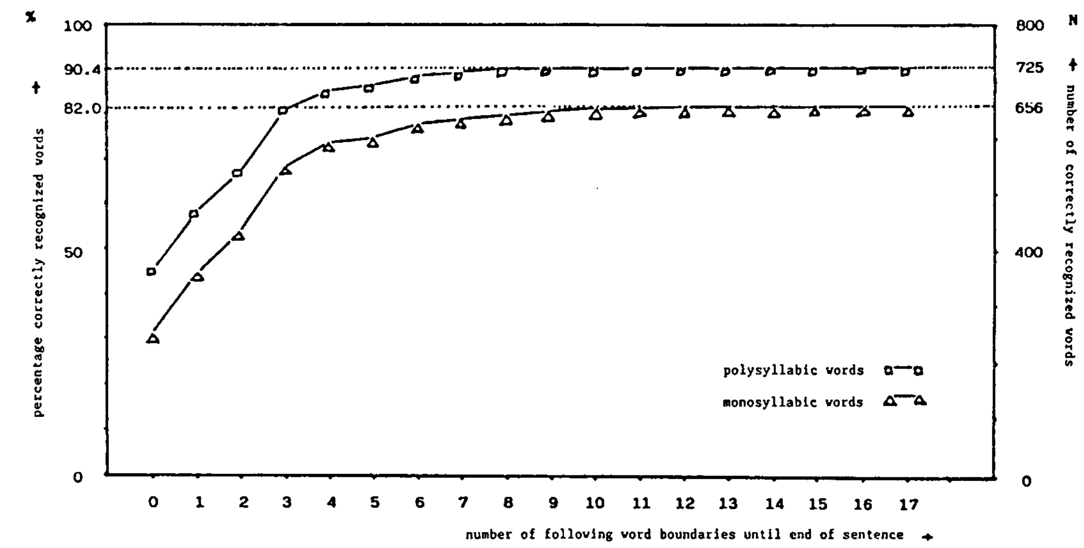


Figure 1. Word recognition of speech synthesized from DIPHONES as a function of the number of word boundaries following a target word. Zero boundaries means: subject heard only part of the target without any following context.

In this figure we have plotted % correctly recognized targets, for synthetic speech only, as a function of the length of the following speech context (expressed in number of words following the target in the audible fragment). Notice, first of all that words synthesized from diphones were recognized less then 40% correct when only their first part (up to and including half of the lexically stressed syllable) is made audible. Even when one foot is added (comprising the integral target as well as at least one other word), recognition is still at 50%. Recognition scores continue to rise as more of the following context is made audible, until 3 complete words have elapsed. The curve then quickly asymptotes when more than 3 words are added to the target.
Context further away than 3 words apparently does not help the listener in finding earlier words that he did not recognize. What has happened when the listener reaches the fourth word? Considering the structure of our stimulus sentences we find that most of the word groups (constituents) contain three words so that the next word is the onset of a new constituent. We argue that later words do not help the listener to recover an earlier unintelligible word across a constituent boundary. This is borne out by the following table which presents percentage content words recognized with or without later context, broken down by word position within the phrase (constituent).

	recognized at 1st partial presentation	recognized after adding one gate	
phrasefinal words 80% (1280)	39% (500)	73% (936)	34% (436)
non-phrasefinal words 20% (320)	35% (111)	79% (253)	44% (142)

Table II. Recognition of synthesized words at different positions in the constituent. N [diphone quality] = 1600. Increased recognition in the case of phrasefinal words is on basis of extra information from a following constituent, in the case of non-phrasefinal words on basis of added information from within the same constituent, $X^2(1)=7.28$ ($p<.01$).

A phrase-penultimate word is recognized on the basis of later context significantly more often than a phrase-final word, $X^2(1)=7.28$ ($p<.01$). We can explain this effect by assuming that transitional probabilities between words are much higher within constituents than across constituent boundaries.

DISCUSSION

Additional context within a constituent seems to enable listeners to recover non-recognized earlier words. We also found that non-phrasefinal words were recovered on the basis of following context more often than phrasefinal words. We take this to be an indication that listeners tend to recognize words in phrases. Therefore, if we are to help the listener recognize words in poor speech quality (synthesized speech), we shall have to mark phrase boundaries with effective prosodic markers.

Acknowledgement:

This research was supported by the Foundation for Linguistic Research, which is funded by the Netherlands Organization for the advancement of pure research, ZWO.

References:

[1] B.A.G. Maassen, "Marking word boundaries to improve the intelligibility of deaf speech", in: Artificial corrections to deaf speech studies in intelligibility, Enschede, Holland, 1985.

[2] S.G. Nooteboom, "The temporal organisation of speech and the proces of spoken word recognition", IPO Annual Progress Report, Eindhoven, Holland, 1983.

STIMULUS CATEGORY, REACTION TIME, AND ORDER EFFECT - AN EXPERIMENT ON PITCH DISCRIMINATION

ANTON BATLINER - LIESELOTTE SCHIEFER

Institut für Phonetik und Sprachliche Kommunikation der Ludwig-Maximilians-Universität München, FRG

ABSTRACT

The "order effect", that causes in a discrimination task the one presentation order to be better discriminated than the reverse order, was tested in the domain of pitch perception with speech and non-speech material as well as with rises and falls. The results showed that (i) rises produce a greater order effect than falls, (ii) non-speech material and rises are better discriminated than speech material and falls, respectively.

INTRODUCTION

The phenomenon of "order effect" (henceforth called OE) has been well known in psychoacoustics since the early thirties, (cf. Stott [7], Zwicker-Feldtkeller [10], Allan-Kristofferson [1]). In the same-different (AX) paradigm, this effect causes the one sequence AB to be discriminated significantly better than the other sequence BA. In psychoacoustic research, this effect has been considered to be an experimental artifact and its influence was eliminated by the following procedure: both orders AB and BA were presented and the mean of the discrimination for both pairs served as criterion for e.g. just noticeable differences, threshold detection etc. cf. [10].
In phonetic research this effect was not dealt with very often (but cf. Repp et al [6], Chuang/Wang [2]). That might be due to the experimental paradigm mostly used in phonetics: in an ABX-task, it cannot show up as clearly as in an AX-task (Repp, 1981 [5]). In our investigations, we used only the AX-paradigm, as it is known [5] that this paradigm is more sensitive than the ABX-paradigm. In several investigations at the Institut für Phonetik in Munich, carried out during the last few years, the OE showed up systematically in studies on speaker recognition (Tillmann/Schiefer/Pompino-Marschall 1984 [9]), tactile discrimination (Tillmann/Piroth 1986 [8]), breathy stops in Hindi (Schiefer, unpublished), German intonation (Batliner, unpublished). In a not yet published paper, we show that the OE is not simply due to the experimental design, and we summarize possible explanations of its origin. In the present paper, we want to address the question of OE from a somewhat different point of view: (i) Does the OE behave differently with speech and non-speech material, i.e. is it a purely psychoacoustic phenomenon, or is there a qualitative difference between speech and non-speech material? (ii) Is there any difference between rises and falls as with regard to the OE?, (iii) What, if any, is the contribution of reaction time to the explanation of the phenomenon? (iv) Is there any difference between the threshold for speech and non-speech material?

MATERIAL

The speech stimulus chosen was 'ja', because the acoustic structure of this stimulus is simple enough so that the factors of interest can be controlled precisely. One of the authors (A.B.) produced several stimuli monotonously in the soundproofed room of the Institute. The stimuli were taped on a Telefunken M15 recorder with a speed of 19 inch per second, digitized on a PDP11/50 with a sample rate of 20 kHz and filtered with a cut off frequency of 8 kHz. For the speech resynthesis of the stimuli a procedure was used where the intensity and the sample points could be defined exactly for each pitch period. The stimulus chosen for the manipulation was segmented into single pitch periods. A logarithmic scale was used for the manipulation of Fo. The stimuli had a constant overall duration of 480 +/-5 ms. The first part containing the fricative, the transition and the first pitch periods of the steady state vowel were left unmanipulated, whereas the remaining pitch periods were subjected to manipulation. Two target stimuli were produced, one falling by one semitone, the other rising by one semitone in its second part. A total of 12teststimuli were derived from the target by increasing the rising contour in six steps of 1/8 tone and decreasing the falling contour analogously in 6 steps of 1/8 tone. These 12 stimuli together with the two target stimuli constituted the body of the speech material. 14 further stimuli were generated, each of which was an exact squarewave analog of the respective speech-stimulus.

PROCEDURE

Four different test-tapes were prepared for each of the subgroups (speech-rises, non-speech-rises, speech-falls, non-speech-falls). In the 'same' condition, each stimulus was paired with itself, resulting in 7 combinations. In the 'different' condition, the target stimulus was paired with each of the other stimuli, the order of presentation being AB as well as BA, resulting in 2*6 combinations. Five repetitions of each of the 19 combinations were taped in randomized order, with an interstimulus interval of 500 ms between the members of a pair. Each pair was followed by a pause of 3500 msec; after 10 pairs a pause of 10 seconds followed. The experiments were run in the speech lab of the Institute with a Revox-trainer and headphones, at a comfortable listening level. Subjects were students that were paid for their participation. They were instructed to compare the two members of a pair, to decide as quickly as possible whether they were different or not, and to press the appropriate button on a box forming part of a digital data collecting device. The responses were collected with a PDP11/03 and prepared for statistic analysis.

RESULTS

Figures 1-4 display the different responses for the orders AB and BA; the number of subjects is given in parenthesis (Fig. 1: speech rises (n=14), Fig. 2: speech falls (n=12), Fig. 3: non-speech rises (n=11), Fig. 4: non-speech falls (n=14)). In all graphs the abscissa displays the difference in tone (1/8 to 6/8), and the ordinate the percent different responses. Generally it turned out that the order AB yields more different responses (i.e. is more prominent) than the reverse order BA. This shows up most clearly for speech rises and non-speech falls, less clearly for non-speech-rises. We are at a loss for any convincing explanation for the unsystematic results for the speech falls.

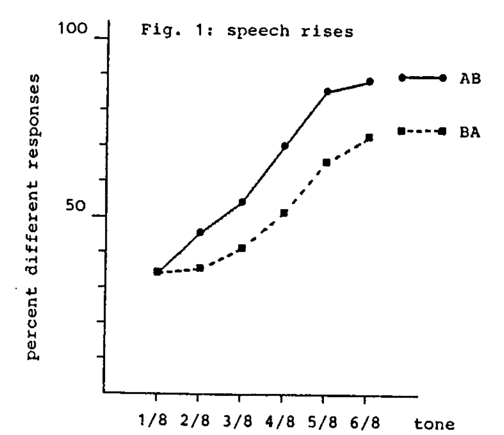


Fig. 1: speech rises

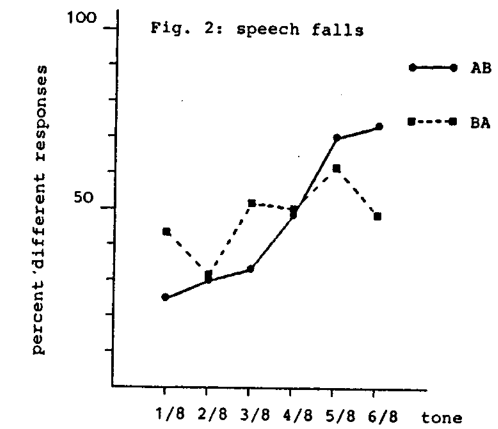


Fig. 2: speech falls

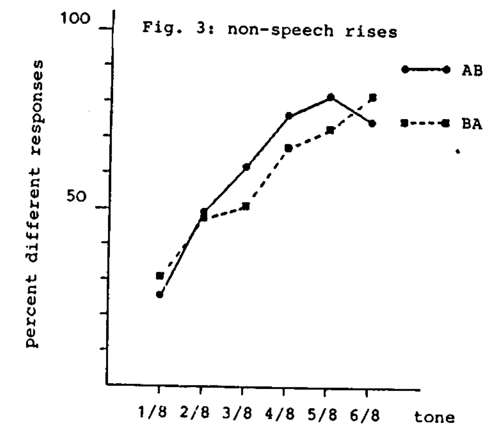


Fig. 3: non-speech rises

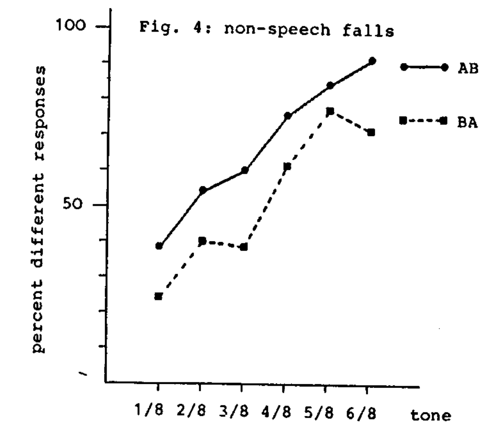


Fig. 4: non-speech falls

A multivariate analysis of variance was applied to the different condition of the four groups together with four factors, two of them being repeated measures (order of presentation AB and BA, difference in tone); the other two, material (speech vs. non-speech) and contour (rise vs. fall) were independent. The level of significance was set to $p < .05$. The necessary assumptions for the multivariate approach were tested with the Cochran and Bartlett tests. Table 1 shows the F-values and level of significance for the effects tested.

<u>Table 1</u>: Statistical results.

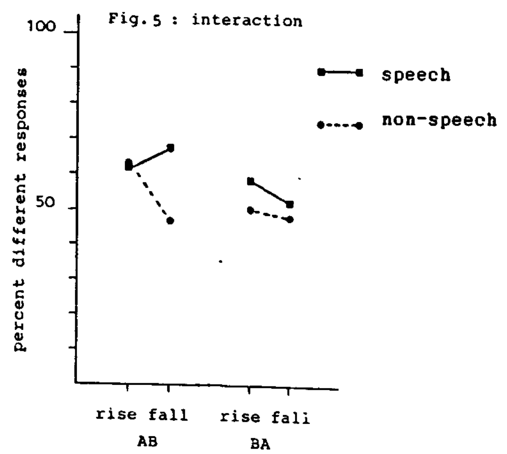
	F	p <
BETWEEN-SUBJECTS (df: 1,47)		
mat. by cont.	1.42	.240
cont.	1.57	.217
mat.	4.22	.046*
ORDER WITHIN SUBJ. (df: 1,47)		
mat. by cont. by ord.	6.95	.011*
cont. by ord.	.03	.860
mat. by ord.	.43	.514
ord.	9.14	.004*
PAIR WITHIN SUBJ. (df: 5,43)		
mat. by cont. by pair	.87	.507
cont. by pair	.60	.694
mat. by pair	1.96	.103
pair	29.09	.001*
ORDER BY PAIR WITHIN SUBJ. (df: 5,43)		
mat. by cont. by ord. by pair	.35	.879
cont. by ord. by pair	2.17	.074
mat. by ord. by pair	1.52	.203
ord. by pair	1.75	.143

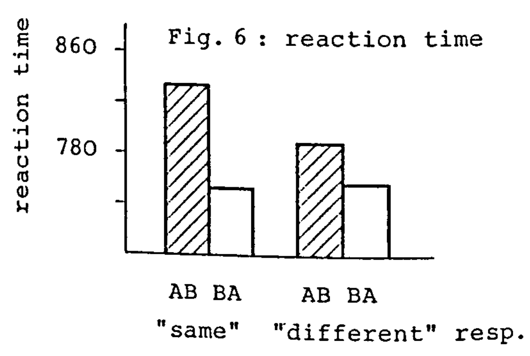
Four of the effects tested turned out to be significant: they are asterisked in Table 1: material, material by contour by order, order, and pair. As there was an interacion between material, contour, and order, the significant main effect of order cannot be interpreted. Therefore, Fig. 5 displays the simple main effects for AB and BA; the interaction shows clearly up in the left part of the figure. Given the presentation order BA (right part of the figure), non-speech stimuli yield more different responses than speech stimuli and rises more than falls. This pattern changes for AB (left part) where no difference between speech and non-speech rises can be observed.
Table 2 shows the intersection of the discrimination function of Figs. 1-4 with the 50% line. We can see, that (i) rises, (ii) non-speech material, and (iii) stimuli in presentation order AB can be better discriminated, than falls, speech material, and BA, respectivly.

<u>Table 2</u>: Points of intersection between the discrimination function and the 50% line.

	speech rises	speech falls	non-speech rises	non-speech falls
AB	2.5	4.08	2.07	1.73
BA	3.86	2.91	2.75	3.5

Fig. 6 displays the mean reaction time (RT) for all four groups taken together. The ordinate shows the RT in ms, the abscissa the 'same/different' responses for the two orders AB and BA. It is obvious that responses to the order AB require longer RTs than those to the order BA, and RTs are shorter for 'different' than for 'same' responses, i.e., hits require less RT than false alarms. (In the 'same' response condition, the difference between the orders AB and BA turned out to be significant, $F(1,1303) = 8.89$, $p < .01$.) These results are comparable to those from the identical pairs, where 'same' responses (i.e. hits) have shorter RTs than 'different' ones.





DISCUSSION

As for material and contour, our results are in agreement with the findings of Klatt [4] and t'Hart [3], who showed that rises are better discriminated than falls and nonspeech better than speech material. The OE turned out to be no purely psychoacoustic phenomenon, as it could be found with the speech and the non-speech material. The present results confirm further our hypothesis, based on earlier findings, that the order AB is better discriminated than the reverse order BA, i.e., stimuli are better discriminated if the stimulus with the greater change in Fo comes last. It doesn't seem to be the height of the offset that is responsable, but the amount of Fo-movement, because otherwise the OE for the falls would favor the order BA and not AB. In the above mentioned paper we will deal with the origin of the OE in detail.

REFERENCES

[1] Allan, L.G. - Kristofferson, A.B.: Psychophysical theories of duration discrimination. Perception & Psychophysics 16: 26-34 (1974)

[2] Chuang, Ch.-K. - Wang, W.S.-Y.: Time-order-error in judgement of prosodic features: pitch, loudness, and duration. Acoust. Soc. Am. 62(S): S48 (1978)

[3] 't Hart, J.: Differential sensitivity to pitch distance, particularly in speech. J. Acoust. Soc. Am. 69: 811-821 (1981)

[4] Klatt, D.H.: Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception. J. Acoust. Soc. Am. 53: 8-16 (1973)

[5] Repp, B.H.: Categorical perception: issues, methods, findings. Haskins Laboratories Status Report on Speech Research SR-70: 99-181 (1982)

[6] Repp, B.H. - Healy, A.F. - Crowder, R.G.: Categories and context in the perception of isolated steady-state vowels. Journal of Experimental Psychology: Human Perception and Performance 5: 129-145 (1979)

[7] Stott, L.H.: Time-order-error in the discrimination of short tonal durations. Journal of Experimental Psychology 18: 741-766 (1935)

[8] Tillmann, H.G. - Piroth, H.G.: An order effect in the discriminability of pulse train sequences. J. Acoust. Soc. Am. 79: S73 (1986)

[9] Tillmann, H.G. - Schiefer, L. - Pompino-Marschall, B.: Categorical perception of speaker identity. Proc. 10th Int. Congr. Phon. Sci., Utrecht, pp. 443-448 (Cinnamison, Dordrecht 1984)

[10] Zwicker, E. - Feldtkeller, R.: Das Ohr als Nachrichtenempfänger. Stuttgart 1967

AN ORDER EFFECT IN PULSE TRAIN DISCRIMINATION AS A CASE OF TIME ORDER ERROR

HANS GEORG PIROTH   HANS G. TILLMANN

Institut für Phonetik und sprachliche Kommunikation
der Universität München
Schellingstr. 3, 8000 München 40, F.R.G.

ABSTRACT

During tactile discrimination tests using the 'System for Electrocutaneous Stimulation SEHR-2', an order effect was revealed: sequences of pulse bursts with constant, small intervals between each two of them were presented in pairs. If these sequences had different intervals, S's discriminated better when the sequence with the larger interval was presented first. In the present investigation the effect was tested systematically in a four-factorial design. A univariate analysis of variance confirmed the order effect, but also showed that discrimination rate is nearly unaffected by variations of the inter-stimulus interval.

INTRODUCTION

In order to establish a system for electrocutaneous speech transmission dynamic pulse train patterns must be designed that serve as 'basic stimuli' for the construction of tactile speech analogues (Piroth 1986 [6]). Those patterns should be readily discriminable and identifiable in a way that is analogous to the perception of natural speech. The 'System for Electrocutaneous Stimulation' SEHR-2 produces current-controlled bipolar electric impulses with freely variable intervals of the form shown in Fig. 1 of Piroth/Tillmann (1984) [7]. This study proved that continua of pulse train sequences that only vary the intervals between pulse bursts can produce a category switch in an identification test, if the threshold of coincidence for successive pulse bursts is crossed along the stimulus continuum: continuous 'movement' on the skin is clearly distinct from 'discrete taps'. Tests with a wide-spread continuum and a 20 ms-step between the neighboring items indicated a threshold at approximately 18 ms. Tillmann/Piroth (1986) [10] revealed an order effect in the discrimination of those stimuli: sequences of nine pulse bursts ('taps') consisting of three impulses each with constant, small intervals (ITI) between each two of them were presented in pairs. Then, if these sequences had different ITIs both shorter than 50 ms, Ss discriminated better when the one with the larger ITI was presented first. The effect could also be reproduced in the auditory mode using sequences of nine short tones and it vanished when, instead of the number of tones, the overall duration of the sequences was kept constant in long tone sequences (2.5 s). Yet, it was not clear, whether the disappearance of the effect was caused by the constancy in overall duration or because there are other parameters governing the perception in a test using sequences with a duration of more than 2s.

The present investigation uses a 2IAX-discrimination experiment to test the dependence of discriminability on the factors 'inter-tap interval' (ITI) 'step'(ITI(B)-ITI(A)), 'order of presentation'(OC), 'inter-stimulus interval'(ISI) in a 5.2.3.3-factorial design so that the nature of the order effect and the role of the threshold for coincidence of successive stimuli can be evaluated based on an interpretation of significant effects revealed by the experiment.

STIMULI

Three pairs of gilded brass electrodes (9 mm in diameter with a minimal distance of 1 mm between the electrodes of a pair) were fastened to the dorsal side of the left forearm. They were arranged linearly so that the distal pair was 3 cm from the wrist, the medial and proximal ones 4 and 8 cm away from the distal one. Sequences of nine 'taps' consisting of three impulses with an impulse width of 200 µs and an inter-impulse onset interval of 2.5 ms each were delivered to the skin. The distal electrode pair received taps 1 - 3, the medial one taps 4 - 6, and the proximal one taps 7 - 9. The sequences differed in the duration of the inter-tap interval (ITI) between the successive taps only.

PROCEDURE

The six Ss participating in the experiment first underwent a calibration procedure to adjust subjective intensity to a mid value between absolute threshold and unpleasentness. Since the 2IAX-test paradigm was used, the tap sequences were arranged in pairs and Ss had to decide whether stimuli were 'same' or 'different' (The arrangement of the factorial parameters is shown in Tab. 1). Each test of the experiment to a third contained pairs of AB-sequences being different with ITI(A) < ITI(B) ('first ordering condition' - 1st OC), to one third BA-pairs ITI(B) > ITI(A) (2nd OC), and to one third AA-pairs with ITI(A)=ITI(A) (3rd OC). ITIs of the sequence with the smaller interval were 5, 10, 15, 20, and 25 ms. Each test consisted of 60 completely randomized pairs (5 ITIs x 3 OCs x 4 repetitions). Each S underwent 6 tests differing in the factors "step", i.e. ITI(A)-ITI(B) (5 ms and 10 ms) and ISI between the sequences of a pair (0.5, 1.0, 1.5 s). The interval between two successive pairs was fixed to 3 s. The tests were presented in different orders to each S so that each possible order of tests was presented once (3 ISIs x 2 steps = 6 tests). Each test was presented twice to yield 8 repetitions per S and combination of factors.

Table 1
The Four-Factorial Design

F1:	ITI = 5, 10, 15, 20, 25 ms	
F2:	Step 1 (5ms)	Step 2 (10ms)
F3:	OC1 ITI : ITI+5	OC1 ITI : ITI+10
	OC2 ITI+5 : ITI	OC2 ITI+10 : ITI
	OC3 ITI : ITI	OC3 ITI : ITI
F4:	ISI = 0.5, 1.0, 1.5 s	

ITI: Inter-tap interval
Step: ITI1-ITI2
OC: Ordering condition
ISI: Inter-stimulus interval

RESULTS AND DISCUSSION

For each S the 8 repetitions were pooled to yield an interval-scaled dependent variable. Fig. 1 shows the data for the three ordering conditions separately for both steps, but pooled over ISIs. Data did not depart from normal distribution or from homogeneity of variance (BartlettBox-Test: $F(89.2007) = 1.05752$, $p = 0.335$). A four-factorial univariate analysis of variance (ITI, Step, OC, ISI) (SPSS [4]) showed significant main effects for all factors for ISI on the 5%-level, for all other factors on the 1%-level. There was no third or higher degree interaction, but Step x OC and OC x ITI interactions were highly significant. On the 5%-level the step x ITI interaction is significant, too (Tab. 2). The significant main effects of step, OC and ITI confirm the results of our previous investigations (Tillmann/Piroth 1986) [10] In varying ISI between the stimuli to be compared we included a new factor in the investigation that is important for the discussion of the order effect. In Tillmann/Piroth 1986 [10] we supposed that the order effect might be explained in terms of the classical "time order error" (TOE) in duration discrimi-

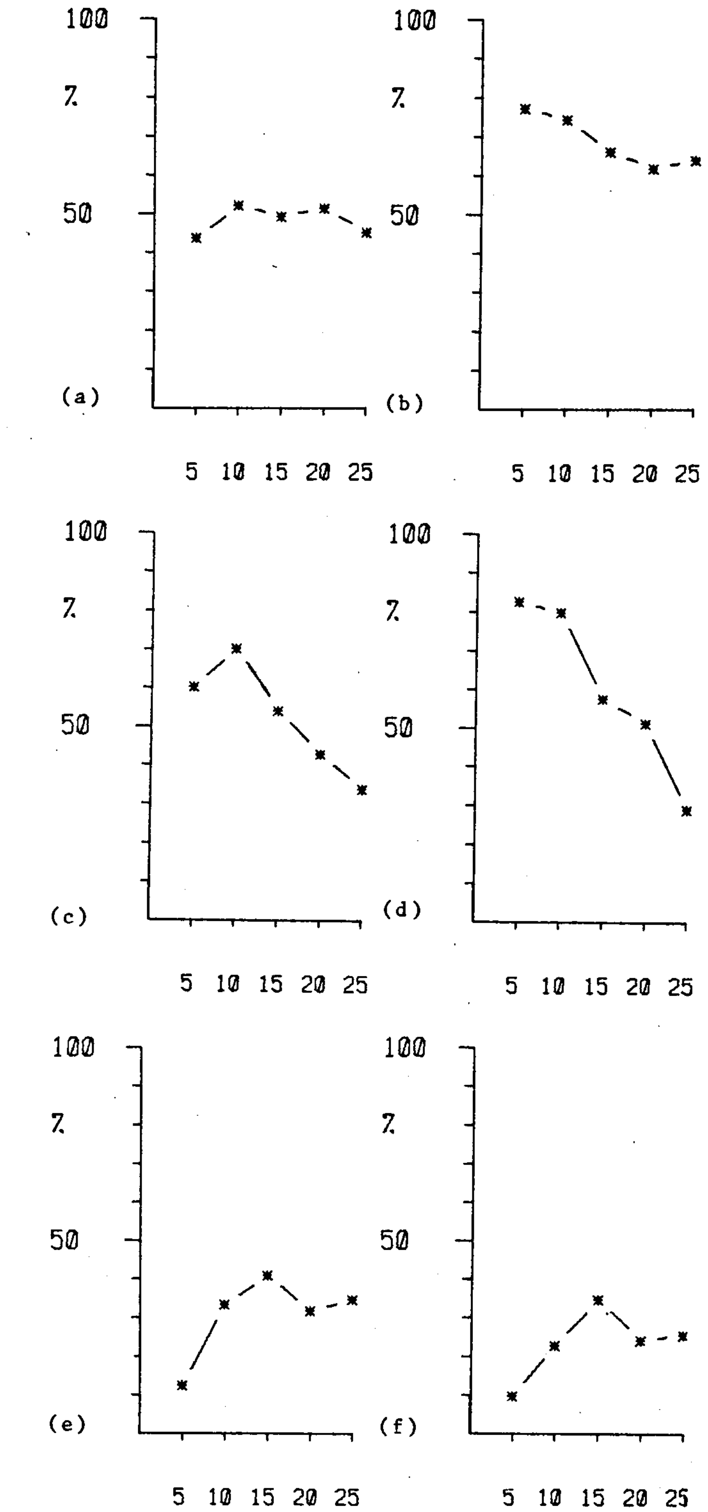


Fig. 1: Percent 'Different'-Responses

(a) 1-step, OC1 (b) 2-step, OC1
(c) 1-step, OC2 (d) 2-step, OC2
(e) 1-step, OC3 (e) 2-step, OC3

Table 2
Main Effects and Interactions

Main Effects:				
F1:	ITI	F(4, 450)	= 12.06453	p<0.01 **
F2:	Step	F(1, 450)	= 18.16458	p<0.01 **
F3:	OC	F(2, 450)	=153.10276	p<0.01 **
F4:	ISI	F(2, 450)	= 3.65926	p<0.05 *
Interactions:				
F1 x F2:		F(4, 450)	= 2.95247	p<0.05 *
F1 x F3:		F(8, 450)	= 15.22227	p<0.01 **
F2 x F3:		F(2, 450)	= 23.82372	p<0.01 **

Caption as in Tab. 1

nation (Stott 1935) [9]. Since the overall duration D of each sequence covaries with ITI (D = 9 x 5.2 ms + 8 x ITI) we may hypothesize that the Ss' judgments in the 2IAX-test are at least partially based on durational differences between the sequences of a pair or ITI differences. According to Allan/Kristofferson 1974 [1] there is another effect that is special in duration discrimination besides TOE: discriminability of durations is not influenced by variation of ISI between the pairs to be compared.
Now, if duration discrimination is an important factor in the Ss' judgments in the present experiment, we should expect that discriminability is unaffected or only slightly influenced by variations of ISI and that the order effect does not interact with ISI. This is the case, since the main effect of ISI is significant only on the 5%-level and since there are no interactions that include the factor ISI (Tab. 2). The main effect of ISI is based on a special contrast between the smallest (0.5 s) and the largest value of ISI (1.5 s) only (ISI1 and ISI3: F(1, 450)=7.01085, p < 0.01). So TOE seems to be at least one factor constituting the order effect.
Further information is provided by the analysis of the simple effects within the significant main effects: Step is only a significant factor within ITI = 5 ms and ITI = 10 ms (F(1, 450) = 23.13967, p < 0.01 and F(1, 450) = 4.02881, p < 0.05). In these cases ITI is in the range of the threshold for coincidence of successive stimuli: Since an ITI of 5 ms and an ITI of 10 ms are both below the threshold it is more likely that sequences in 2-step pairs belong to different categories ("continous movement" and "discrete taps"). This interpretation is confirmed by an analysis of the special contrasts: only the contrast ITI = 5 ms vs. ITI = 10 ms is not significant (ITI1 and ITI2: F(1, 450) = 0.05887, p = 0.808; ITI2 and ITI3: F(1, 450) = 27.68914, p < 0.01; ITI3 and ITI4: F(1, 450) = 16.42786, p < 0.01; ITI4 and ITI5: F(1, 450) = 4.08225, p < 0.05).
Step is highly significant only in OC1 and OC2 (OC1 : F(1, 450) = 51.73833, p < 0.01; OC2 : F(1, 450) = 7.53802, p < 0.01). OC3 consisted of pairs of physically equal sequences. Nevertheless, step is significant on the 5%-level in OC3, too (F(1, 450) = 6.53567, p < 0.05). This fact seems to reveal a contextual effect in discriminability: the sameness-criterion used by the Ss' (Pollack/Pisoni 1971 [8]) seems to depend partially on the overall inventory of sequences presented during the test run. OC is highly significant for both steps and all ITIs, but ITI is significant only in OC2 and OC3 (OC2 : F(4, 450) = 31.75895, p < 0.01; OC3: F(4, 450) = 9.54535, p < 0.01). This can be explained by concidering the form of the discrimination curves: in OC2, discriminability decays with increasing ITI, so that a significant effect arises, in OC1 discriminability is bad and remaines nearly constant along the ITI-continuum. In OC3 the effect is due to a significant variation of 'different'-answers along the continuum that yields a peak when ITI is 15 ms.
The analysis of the special contrasts within the order effect shows that there is a significant difference between OC1 and OC2 as well as between OC1 and OC3 (OC1 and OC2: F(1, 450) = 60.04474, p < 0.01; OC1 and OC3: F(1, 450) = 245.16078, p < 0.01). Now, an order effect in the discriminability of stimuli in a 2IAX-discrimination test can arise if perceptual equality of the pairs departs from physical equality. Tillmann/Piroth 1986 [10] argue that in pairs of sequences of the kind used in the present investigation the second sequence has to be longer than the first to be sensed as being 'same'. From another point of view, this means that the physically equal sequences of the AA-pairs in OC3 are not perceptually equal. Additionally, the sequence with ITI = 15 ms is in the range of the threshold for coincidence of successive stimuli. According to the theory of categorical perception, discriminability increases near category bounderies (e.g. Liberman et al. 1967 [3]). Since 'continuous movement' and 'discrete taps' are established categories in natural cases of tactile perception, we can suggest that there is a peak in discriminability, if two stimuli of different categories are presented in a pair. According to our assumption on perceptual equality the second sequence of an AA-pair consisting of two sequences with ITI = 15 ms will be sensed to be faster than the first. Since ITI = 5 ms and ITI = 10 ms are clearly below the coincidence threshold, the first sequence of the 15ms:15ms-pair is possibly sensed to be above and the second to be below this threshold. This may explain the peak in discriminability found for physically equal pairs of sequences with ITI = 15 ms.
The discussion of order effects is not uncommon throughout literature on normal speech perception, too (e.g. Ohde/Sharf 1977 [5], Uselding 1977 [11]).
Even the order effect in duration discrimination (TOE) may be found in speech data. Thus, Lehiste 1973 [2] introduced the notion of 'final lengthening' to describe the phenomenon that in syllable sequences syllable duration increases at the end of the sequence. Now, our presentation of tap sequences in 2IAX-pairs (which are at least minimal sequences of two stimuli) caused the effect that the second member of the sequences had to be longer to be perceived as being as long as the first. As mentioned, it was possible to reproduce with acoustic stimuli. So, one might suppose that a similar effect concerning the perception of syllable duration is compensated by 'final lengthening': since the last syllables are physically longer in duration they might be sensed as being as long as the proceeding ones.

REFERENCES

[1] L.G. Allan, A.B. Kristofferson, "Psychophysical Theories of Duration Discrimination", Perc. & Psychophys. 16, 1974, 26-34.

[2] I. Lehiste, "Rhythmic Units and Syntactic Units in Production and Perception", J. Acoust. Soc. Am. 54, 1973, 1228-1234.

[3] A.M. Liberman, F.S. Cooper, D.P. Shankweiler, M. Studdert-Kennedy, "Perception of the Speech Code", Psychol. Rev. 74, 1967, 431-461.

[4] N.H. Nie et al., "SPSS Statistical Package for the Social Sciences. New York etc. 1975.

[5] R.N. Ohde, D.J. Sharf, "Order Effect of Acoustic Segments of VC and CV Syllables on Stop and Vowel Identification", J. Sp. Hear. Res. 20, 1977, 543-554.

[6] H.G. Piroth, "Electrocutaneous Syllable Recognition Using Quasi-articulatory Coding of Stimulus Patterns" (Abstr.), J. Acoust. Soc. Am. 79, 1986, S73.

[7] H.G. Piroth, H.G. Tillmann, "On the Possibility of Tactile Categorical Perception", M.P.R. v.d. Broecke, A. Cohen, Proc. 10th ICPhS. Dordrecht 1984, 764-768.

[8] I. Pollack, D. Pisoni, "On the Comparison Between Identification and Discrimination Tests in Speech Perception", Psychon. Sci. 24, 1971, 299-300.

[9] L.H. Stott, "Time-order Errors in the Discrimination of Short Tonal Duration", J. Exp. Psychol. 18, 1935, 741-766.

[10] H.G. Tillmann, H.G. Piroth, "An Order Effect in the Discriminability of Pulse Train Sequences" (Abstr.), J. Acoust. Soc. Am. 79, 1986, S73.

[11] D.K. Uselding, "A Temporal Order Effect in Voice Onset Time Discrimination", Lang. & Speech 20, 1977, 366-377.

ACKNOWLEDGEMENTS

This investigation was supported by the German Research Council (DFG-grant Ti 69/25).
Statistical advice was given by Dr. Alexander Yassouridis, Departement of Biostatistics, Max-Planck-Institute for Psychiatry, Munich F.R.G.

PERCEPTION OF RHYTHM AND ITS ROLE IN THE PROCESS OF LANGUAGE ACQUISITION

MORIO KOHNO

Kobe City University of Foreign Studies
Kobe, Japan 654

ABSTRACT

The mode of table tapping, for example, had a striking similarity with the mode of utterance in Japanese in terms of speed and of interstress (-beat) intervals, which eventually concentrated in the area of 400-1,000 msec. in both modes. And more interestingly, the subjects who tapped quickly outstripped the slow tappers of young ages in the ability of learning a new foreign language, even if they are very old, i.e., 80 years old.

Other experiments revealed that, when interstress intervals go out of the central rhythmic area (400-1,000 msec.), intelligibility of the speech abruptly falls in the case of initial stage learners of a foreign language. These and some other evidence seem to throw doubts on the widely-spread ideas of gradual perception of rhythm and of the critical period in language acquisition.

INTRODUCTION

Rhythm has been acknowledged to be one of the important, probably universal, principles in spoken language, functioning both as an organizing factor in speech articulation [1], and a guiding principle in the perception of speech [2]. If there should be a possibility that rhythm might be innately acquired or a universal phenomenon as some phoneticians suggest [3], it would possibly play a very important role in the process of understanding sentences and discourses, and even in the process of language acquisition, both of which require far more complicated strategies and cognitive actions than the perception of sounds, or isolated words.

EXPERIMENT I

An English short story (106 words) was composed by the use of the words and sentences which had already learned by the subjects -- Japanese high school students (9th grade) who were learning English (n.=120). The content of the story was original, however. This story was read and recorded by an English speaker at three reading speeds: normal, fast and slow. The normal speed material was then compressed mechanically by Speech Compressor (HITACHI TSC-8800) (Machine fast). These four modes of the material had the following acoustic characteristics as a result of measurement by Visicorder and Electro-oscillograph (Yokogawa, Type 2901, amplifier 3125):

Reading speeds: normal=124.8 words/minute, slow=74.9, fast=198.7, machine-fast=185.8

Frequency of juncture pauses: normal=17, slow=35, fast=14, machine-fast=16

Total amount of juncture pauses: normal=14,098 msec., slow=32,693, fast=7,209, machine-fast=11,071

Interstress intervals (means): normal=574 msec. (S.D.=154.8), slow=568 (S.D.=179.2), fast=446 (S.D.=90.6), machine-fast=359 (S.D.=144.6)

(The phonetic stoppages whose duration was less than 220 msec. were not included in the data on pauses, because this kind of discontinuation often occurred when plosives were made.)

Intelligibility of the four modes of the material was got by asking questions of the subjects in Japanese about the content of the story, as follows:

normal > fast, $p < 0.01$, ($t=3.40$)
normal > machine-fast, $p < 0.01$, ($t=6.18$)
slow > normal, $p < 0.05$, ($t=2.52$)
slow > fast, $p < 0.01$, ($t=5.92$)
slow > machine-fast, $p < 0.01$, ($t=8.70$)
fast > machine-fast, $p < 0.05$, ($t=2.78$)
(A > B in the table means A mode gave more intelligibility than B mode to the subjects who were homogeneous in ability of English.)

Obviously no other factors than interstress intervals (rhythm) can interpret the above-mentioned facts. The speed of utterance and the pausing conditions can never explain the phenomenon of fast > machine-fast, but the interstress intervals can -- they are less than the central rhythmic area (400-700 msec.) suggested by Allen (1975) [4], in the machine-fast.

The author's previous paper [5] shows that pauses help listeners' cognitive processing if normally placed, but too many pauses which were placed at the ends of every word, for example, hinders listening comprehension. The paper also shows that slowly drawled pronunciation which is too accurately articulated has the same negative effect. The reason for this phenomenon can be explained by the fact that, while the interstress intervals of the former material were 574.2 msec., the ones of the latter two were 1845 and 2643 respectively, both of which are off from the central rhythmic area.

Several additional experiences were held in the similar way as Experiment I, changing the conditions of interstress intervals variously and revealed that intelligibility falls when they go out of the area of about 400-1,000 msec. It may be said in this connection that interstress intervals of the first cry of a newborn baby was around 756 msec. (means).

EXPERIMENT II

Subjects are Japanese with age variety from 18 through 79 years old (n.=30). They were asked to tap a table at a tempo they feel most normal, and then to read "The North Wind and the Sun" by Aesop at normal speed in Japanese. The result is shown at Table 1:

Table 1

Subjects	Tapping (f/m)	Reading speed (w/m)	Age
1	107	211	18
2	105	208	18
3	102	198	22
4	95	192	20
5	100	185	49
6	96	182	43
7	96	180	23
8	90	177	24
9	86	175	55
10	78	173	16
11	70	170	35
12	77	170	14
13	47	170	49
14	75	167	41
15	73	165	57
16	60	164	59
17	45	160	35
18	70	155	17
19	66	148	71
20	90	145	23
21	60	143	39
22	60	143	37
23	60	140	62
24	56	140	40
25	45	136	65
26	49	135	65
27	86	132	23
28	65	130	55
29	48	123	74
30	43	111	78

The rank correlation between the tap frequency and the reading speed is very high ($rs=0.78$, $t=4.2$, $p < 0.0001$), and those between tapping and age, reading speed and age are $rs=0.61$, $t=3.28$, $p < 0.02$ and $rs=0.57$, $t=3.06$, $p < 0.004$, respectively.

Slow, rapid and normal speed tappers, 7 in total, were then chosen from the subjects, and they were required to learn Spanish, which was never learned by any subjects in the past. The content of learning (testing) was. 1) to repeat some Spanish words and sentences without models after having listened to them by a tape recorder five times. 2) to find some grammatical rules heuristically after having listened to some sets of words and sentences. 3) to repeat phonemically minimum pair words after having listened to them five times. All the responses were tape-recorded and scored by two teachers of Spanish. (coefficient of objectivity=0.82)

Table 2

Subjects	Spanish test score	Tapping f/m	Reading condition: interstress interval msec.	Reading condition: S.D.	Age
1	108	110	412	15	18
5	102	104	453	15	49
8	98	96	525	21	24
19	77	66	755	21	71
28	59	65	870	30	55
29	51	48	931	31	74
30	57	43	905	30	78
EX	73	75	692	21	79

Relation between the Rhythmic Actions and Language Acquisition

Table 2 shows the very high correlation between the frequency of tapping and test scores, but between age and test score, the correlation is not so high. In order to confirm the relation of neuro muscular ability for rhythmic action with competence for language acquisition, a male subject, EX, who is still quick in action in spite of being 80 years old was asked to join the experiment, which brought forth the result described in the bottom of Table 2 -- his tapping is very smooth and his test score is also high.

We can conclude rhythm is unexpectedly crucial in the process of acquisition, closely connected with human beings' motor actions.

References

[1] J. Martin, "Rhythmic (hierarchical) vs. serial structure in speech and other behavior" Psychological Review 79: 487-509, 1976.
[2] W. Marslen-Wilson and L. Tyler, "The temporal structure of spoken language understanding" Cognition 8: 1-71, 1980.
[3] M. Studdert-Kennedy, "Speech perception" Proceedings of the Ninth International Congress of Phonetic Sciences II: 59-81, University of Copenhagen, Denmark, 1979.
[3] I. Lehiste, "Isochrony reconsidered" Journal of Phonetics 5: 253-263, 1972.
[4] G.D. Allen, "Speech rhythm: its relation to performance universals and articulatory timing" Journal of Phonetics 1: 219-237, 1975.
[5] M. Kohno, "The effects of pausing on listening comprehension: some psycholinguistic experiments in the case of Japanese learners of English" in T. Konishi (ed.), Studies in Grammar and Language 392-405, Kenkyu-sha Tokyo, 1981.

THE DESIGN OF A SPEECH ANALYSIS WORKSTATION

JOHN M. CRUMP

Kay Elemetrics Corp
12 Maple Ave.
Pine Brook, NJ 07058 USA

ABSTRACT

The development of a speech analysis workstation is presented. The problems and challenges in acoustically analyzing speech signals are discussed. A system was developed to provide the digital acquisition and analysis of speech with all of the features typically required in acoustic phonetic research.

INTRODUCTION

Speech has been acoustically analyzed by a wide assortment of instruments including oscilloscopes, spectrographs, and numerous computer based systems. Typically a computer system requires a number of peripherals to analyze speech. These peripherals may include input modules with A/D and anti-aliasing filter, graphic boards and special printers. High speed array processors or special digital signal processing boards may be added to boost processing speed. Software to analyze the stored signal is typically purchased commercially or developed by researchers.

The recent availability of general purpose digital signal processing chips, inexpensive digital memories and personal computers has provided the technical capabilities for the development of a powerful workstation designed for the analysis of speech. A system can now be developed with the advantages of a spectrograph (e.g. Sona-Graph and SSD), an oscillograph (e.g. Visicorder), a feature extractor (e.g. Visi-Pitch), and a general purpose computer (e.g. VAX with DSP software).

DEFINING A SPEECH WORKSTATION

Before the development of a speech analysis workstation is started, it is important that the analysis requirements of the users are clearly understood. Speech is analyzed by many different professionals for many different reasons. A phonetician may have different needs than an speech language pathologist. Any workstation designed for speech analysis must take these different requirements under consideration. The common elements for most speech analysis are reviewed as follows:

Input

The aliasing portion of a signal must be filtered before the signal is digitally stored. Low-pass filtering is the process of eliminating the high frequency components which will create spurious spectra in the analysis. Providing adequate anti-aliasing filters is a difficult, and often overlooked, problem especially if the user changes sampling rates to perform different analysis tasks. For example, the analysis of vocal behavior (e.g. perturbation measurements) requires very high sampling rates to achieve high timing accuracy. Sampling rates as high as 50-100kHz may be required. Anti-aliasing filters at these sampling rates are quite different from filters at slower sampling rates.

Sampling frequency must be variable and should exceed the 50 kHz sampling rate required in some applications.

Psycholinguistic experiments and phonetic transcription require a system which can store and playback speech at high fidelity. High fidelity playback requires high sampling rates. If the workstation is to be used to acquire and define a phonetic library the speech signal requires a deep dynamic range and excellent frequency response. Dynamic range should be above 70dB and sampling rates above 50kHz. The speech signal storage should be sufficient to store at least one paragraph of speech sampled at high rates.

All of the above requirements are very important because there is a general requirement for instrumentation to simply acquire, filter, amplify/attenuate, A/D, D/A and buffer speech signals for input to computers for further analysis. A speech workstation should be able to excel in this limited but important function.

From the requirements explained above the following criteria for input and signal storage were developed:

Sampling rates: Variable with samples up to 80kHz

Dynamic Range: 12 bits or >72 dB

Low pass filters: Automatic with sampling selection, 120 dB/octave, preferrably digital filters.

Signal storage: At least 40 seconds sampled at 20kHz. This requires 2 Mbytes of memory.

Displays

Graphically, speech has tradionally been displayed as a waveform, a spectrogram, a power spectrum (frequency vs. power) or as tracings of speech parameters. A speech workstation should be able to present these four standard displays clearly and crisply. Speech analysis also typically requires timing and frequency measurements. Various feature extraction techniques such as LPC analysis has also proven itself a useful tool. Integrating these various approaches in the analysis of speech would be especially useful. For example it would be useful to superimpose color LPC extracted formant values on a wide band grey scale spectrogram. Depending on the analysis task it would also be desireable to be able to rapidly switch analysis formats to find the type of display most revealing of the characteristic under investigation.

A workstation should allow a wide range of display options which can be quickly performed (less than 2 seconds). This will help users quickly re-analyze the stored data to find the most revealing display of the aspect of interest. Time resolution of waveform displays must facilitate the measurement phenomenom of both very short and long duration. Timing accuracy should be as fine as each data point of memory for resolution of 0.01 milliseconds. Spectrograms must include a selection of analysis filters for the fine time and frequency resolution frequired for the effective formant display of low and high pitch voices.

Real Time Performance

Real time analysis is valuable for a number of reasons, some obvious and others not so obvious. The faster the analysis is performed the less waiting for the user. If the user can quickly re-analyze data he or she is more likely to explore various analysis modes to find the most revealing method. In any clinical setting real time analysis is usually a requirement.

The other advantage of real time analysis is that the data can be monitored during input and analysis. Systems, which batch analyze data, require the user to first store data and then analyze. Speech is such a dynamic signal that unless the input can be monitored during input it is very difficult to acquire the signal without overloading during transient peaks or underutilizing the full dynamic range. One solution is to use input systems with very deep dynamic range (>90dB) which require 16 bit A/D and extremely good low noise input circuits and anti-aliasing filters. These systems are very expensive.

For many applications it is important to monitor the analysis in order to select the correct data for analysis. For example if the researcher is investigating an acoustic phenomenon which is clearly displayed spectrographically, but is difficult to hear, real time capability allows the user to scan the input speech signal to select the appropriate segment.

Some systems will analyze in real time, but can not simultaneously store the speech signal. This is obviously undesireable because the user must re-enter the signal to re-analyze. A true real time system must be able to simultaneously low pass filter, acquire, store to memory, analyze and display in real time.

Graphic Resolution

As mentioned above the graphic displays are an important component in any speech analysis workstation. High resolution graphic displays are technically difficult. Typical microcomputers video graphic standards fare not good enough to replicate the display resolution of even 1950 style hard copy spectrograph. The selection of grey scales available are insufficient to display spectrograms. The fine timing and frequency measurements require a more robust display standard with more than 32 shades of grey for each element and a display resolution of at least 640(H) x 480 (V). Hard copy resolution must match the standard set by the commonly available hard copy spectrographs. A color display would also be useful to display speech parameters (such as LPC extracted formant frequencies) and grey scale spectrograms simultaneously. Color is also required when multiple traces are displayed.

Interface to Computers

A speech workstation should be able to operate inside a microcomputer, or be easily interfaced to microcomputers. For a number of reasons discussed in more detail in another section of this article currently available microcomputers can not become practical speech workstations. Despite these limitations inexpensive microcomputers can serve valuable functions if interfaced to a speech workstation. The availability of inexpensive file management, data storage and software complement the analysis and display power of a speech workstation. An interface to these microcomputers should be very fast to facilitate rapid exchange of data files and to increase the utility of the speech workstation as a data acquisition peripheral.

Programmability

The rapid advances in digital signal processing of speech necessitate that a speech workstation can be updated to apply new algorithms to speech analysis. Often users are only interested in a single speech analysis measurement and may require adjustments to currently available programs to best extract this information. It would be desirable for the user to be able to change programs and a requirement that the vendor can upgrade without using software rather than hardware replacement.

User Friendly

A speech analysis system will often be used by speech scientists, speech language pathologist and phoneticians who may not be instrument oriented or computer specialists. They also may only perform acoustic analysis infrequently in their work. In this working enviroment, it is important that a speech workstation is easy to use. The system should be menu driven and methods of analysis/display should be electronically storable and retrievable so that users can repeat analysis methodology exactly.

In a teaching enviroment acoustic analysis tools are often used to teach students about acoustics. It would be useful for a workstation to be designed to facilitate this task by storing precisely repeatable acoustic analysis experiments.

Dual channel Capablility

Speech is often investigated in conjunction with other physiologic signals. A speech workstation should be able to operate in dual channel mode to analyze electroglottograph, airflow, accelerometer and other signals of interest in conjunction with the speech signal.

Affordability

Price and performance have obvious tradeoffs in any development but a speech workstation can not be beyond the reach of most speech scientist no matter how wonderful the product is.

EXPLORING THE AVAILABLE TECHNOLOGY

Once the outline of the features and specifications were established the commercially available technology was investigated to determine the best approach to accomplish the design criterion. One approach which was considered in detail was the packaging of the hardware/software for this workstation inside a standard microcomputer. In this configuration the hardware would plug into the backplane of the DEC Q- Bus, the IBM-PC bus or directly connect to a high speed port of other computers. DEC , IBM-PC ATs, Amigas, Apollo, Sun, Masscomp, MacIntosh and others were evaluated.

Incorporating the workstation in these comon computers was rejected for technical and/or cost considerations. The widely available inexpensive computers (IBM-PC, Amiga, MacIntosh etc.) were not powerful enough even with added hardware. The technical limitations of inexpensive microcomputers to perform as a speech workstation are as follows:

1. The bus of microcomputers has a very limited bandwidth and it can not, therefore, acquire signals at the sampling rates required for many speech analysis tasks.

2. The bus and DMA capabilities of microcomputers do not allow the simultaneous transfer of data from input board to memory, input board to analysis module, analysis module to display memory. It can not, even with the addition of graphics, input and digital signal processing boards do true real time acquisition, analysis and display.

3. Most computers have insufficient memory available for signal storage. As noted previously at least 2Mbytes of signal stoage are required in addition to 512K bytes of digital signal analysis work space and 384kK bytes of display memory.

4. The digital signal processing speed is at least 100 to 200 times too slow for real time analysis. Accelerator boards can be added but the speed is still insufficient for a robust system.

5. The highest standard graphic standards on microcomputers are not able to display spectrograms with enough resolution in time (horizontal), or sufficient grey scale. Many computers restrict the user to specifc color selections because the video controller can only turn on or off each RBG output guns. This restriction does not allow the subtle variation of hue or grey scale necessary in some applications.

The more powerful systems are costly and not widely or consistently available for many potential users. Even these more powerful systems (VAX etc) are too slow for the real time digital signal processing required. Array processors would need to be added to achieve real time performance and, in some cases, the system architecture can not transfer data blocks at the required rates.

These technical and cost considerations aside, it must still be emphasized that it is important to have high speed interfacing between the standalone speech workstation and the widely available IBM-PC type microcomputer and VAX minicomputers. High speed interfacings eliminates the need for the workstation to include its own disk drives and allows access to available DSP software and previously digitized data.

DEVELOPING THE WORKSTATION

The result of the exploration has led to a standalone system based on a common microprocessor, powerful digital signal processing integrated circuits, high resolution graphic displays and high speed DMA capabilities. The digital signal processing chip selected was the 32020 from TI (Texas Instruments). Two 32020s are used to further increase the processing speed to ten million instructions per second. The 32020 are capable of many parallel operations and include a fast single-instruction multiply operation. These features are extremely useful because of the repetitive nature of the instructions and the many multiplications required in digital signal processing. These features combine to provide digital signal processing speeds equivalent to over 50 million instruction per second in a general purpose computer. These chips were also selected because of the upward migration path TI has produced with the 320C20 and 32030.

Two separate buses for data acquisition and analysis were used. This "extra bus" and special high speed DMA chips were used to facilitate high speed data transfer between the different system modules (A/D to memory, memory to DSP circuits, DSP circuits to graphic circuits and DSP circuits to printer). These DMA chips allow a 4 Mbyte/sec transfer rate. The system management is performed by a Motorola 68000 and the system architecture has been defined to include up to 8 Mbytes of RAM and 2 Mbytes of PROM.

The graphic resolution required for both the real time display monitor and hardcopy were the most difficult to achieve. The system was designed with a graphics controller, high speed video DRAMs and a special monitor to provide graphic resolutions of 640 x 480 with 256 values of color and/or grey scale for each pixel. The sytem allows simultaneous grey scale and color displays because the monitor and video driver are capable of both analog and digital display. Because the extensive graphics routines required in a speech workstation can not processed quickly through the CPU the graphics hardware was designed to perform most of the graphics displays without CPU intervention.

The hardcopy print capability is based on a new thermal printer and this print quality matches the quality of Kay sonagrams™ which have become the standard for spectrographic display. The printer produces true (not imitated with a collection of dots turned on or off) grey scale at 120 dpi.

The system requires multiple processing modules to achieve the speed and performance required. The relatively slow CPU is relieved of virtually all of the processing, except for controlling the other modules.

The system meets all of the criteria set above for a speech workstation. It can not be programmed by a novice and is, therefore, limited to the programs available from Kay or programs developed by programmers familiar with the TI 320 code. There are over 320 design teams working with this chip according to TI. How many are working in the speech field is not known but the TI320 family represents over 65% of the digital signal processing chips sold in 1986. It has become a standard for digital signal processing development and there are numerous plug-in boards for computers designed for 320 code development. Kay has developed a series of programs to implement all of the features discussed in the the section "DEFINING A SPEECH WORKSTATION". Along with the development of numerous speech analysis programs continuing at Kay other groups, including the University of Victoria's CSTR (Centre for Speech Technology Research), are working on LPC analysis / modification/ synthesis programs. Kay will commercialize the programs developed by CSTR. The system has all of the programs stored on a large PROM board to facilitate updates as the science of digital signal processing develops.

When interfaced to computers the speech workstation can also be used for input, speech selection and buffering, display and grey scale printing. Users can then use the programming tools available on their computer for other digital signal processing or file management programs.

SUMMARY

The system succeeds in meeting the design criterion for a general purpose standalone workstation. State of the art technology and multiple processing modules were required to meet this criteria. To facilitate its utility as a peripheral to common computers such as the IBM-PC and DEC VAX, software is being written to exchange data and allow these computers to easily use the powerful graphics, data acquistion and digital signal processing capabilities of the standalone speech workstation.

THE SPEECH LAB

Jan Sedivy, Jan Uhlir

Czech Technical University, Prague
Department of Circuit Theory

Introduction.

This contribution describes a simple programming environment called Speech Lab (SL). The system is designed for personal computers operating under CP/M-80 or MS-DOS.

The SL was designed at the Department of Circuit Theory of Czech Technical University in Prague for the speech processing. The scope of task being solved is very large. It includes the basic signal processing algorithms, wave-shape coding methodes, LPC analysis etc. Recently a simple knowledge base was added to support works in speech synthesis.

SL is used in all speech oriented works in our department. It is used by students, by after graduate students and by staff of the department. It enables easy way of data and program exchange.

The SL structure.

The SL is structured programing environment. It consists of the following building blocks:

- User interface
- Data Acquisition system
- Data Processing Package
- Graphics Package

The SL is controlled trough the User Interface. It is made up from powerful commands. Commands can be divided in following groups, which correspond with the building blocks:

- Database Commands
- Data Acquisition Commands
- Data Processing
- Graphics Commands
- External Procedures
- Help Menu System

The first set of commands are database commands. They are used to display the database records, to retrieve, erase and update records. Other commands are devoted to get and put the data on the disk. Special command serves to import and export ASCII files.

The Data Acquisition Commands cooperate with the D/A and D/A convertors. They assure a fast data acquisition and the immediate check of processed speech. A simple command "SPEAK" is very useful in many speech applications.

For simple data processing are available many commands. Some of them are build-in and some are external commands. They are used for signal processing.

A very important feature of the SL is the simple ability to make up own procedures. The whole SL is written in TURBO Pascal [2]. This implementation of Pascal programing language is very popular in the PC compatibles and CP/M-80 world. The user created procedures written in TURBO Pascal can be called from SL with single command. They can use the same data as SL. It is an easy task to go back to SL after processing the data in a common block. This is accomplished using the TURBO Pascal command chain. In this way the user can use all the SL commands and create his own procedures. The burden of all data housekeeping and many processing is minimized.

User interface is equipped with large help menu system. This help can be called any time.

The SL uses a 20k or 32k long buffer for the data storage depending on the particular implementation. This buffer is used as above mentioned common block which is used by both SL and user-written procedures. To simplify the orientation in data, the buffer is divided in particular number of 128 bytes long sectors. Every byte in any sector is user accessible. The same structure of the data pointer is used in all commands. The data pointer is composed from two numbers. The first number is used for a sector and the second for a byte in the sector. We found this type of pointers useful for the sort of applications the SL was designed for.

The data structure.

The data in the buffer are speech samples or any other type of data. For example the LPC coefficients, spectrum, histogram etc. This data can be processed and then graphically displayed on the screen. There are available different graph shapes for the spectrum, histogram or the speech time series. User can chose to display the data in any shape. The graphs can be print out. Simultaneous display of different records is possible.

The file structure.

The SL record consists from two files a data file and a dictionary file. The dictionary file carries the information about the data like the sampling frequency, the date of creating the record, remarks etc. The data file contains pure data. Both files are automatically maintained using the SL commands. The dictionary file is short and it is presented in the SL on line to speed up the directory operation. The data file is updated after any changes has taken place.

A/D D/A converters.

The SL system enables an easy installation of A/D and D/A drivers. The drivers must be written in assembly language and must be patched in the SL. The SL supports any sampling frequency for data acquisition. We usually use double sampling frequency then required, to decrease the analog filter requirements. The final data are the digitally filtered with FIR filter. Simultaneously is the signal decimated in frequency. In this way the linear phase response can be preserved.

Graphics.

The SL uses Graphical Commands to display the contents of the BUFFER on the screen. The way how it is displayed depends on the hardware abilities. The graphical system is a separate part of the SL. It is written for different graphical systems. The CP/M versions are not so rich in graphical abilities compared to MS-DOS version. Both versions give the same type of the printer output.

The user can chose a proper shape of the graph different types. Shapes are designed to display a speech wave-shape, the power spectra, the histogram, the autocorrelation function etc. In the graphical part of the SL are also included commands to display a difference of two signals.

Graphs can be send to the printer or plotter.

Data processing.

The SL has build-in the most important signal processing procedures. They can be applied to active buffer. These procedures are:

- Histogram
- Autocorrelation
- Windowing
- FFT

External procedures are in the processing library. It contains a large set of algorithms for spectral and LPC analysis.

Different algorithms for speech wave-shape coding can be called. For example different types of PCM and DPCM quantization, adaptive versions of PCM and DPCM etc.

We also developed a package of clustering algorithms for the purpose of vector quantization of descriptors or wave-shape of speech.

For the purpose of synthesis by rule we developed a special package with a simple knowledge database. It is used for development of synthesizers with limited number of words and high quality of speech. The knowledge database simplifies the construction of similar sounds.

User written-subroutines.

A very important feature of the SL is the ability to include a user-written subroutine and to use the facilities of the system too. The whole system is written in Turbo PASCAL. The data buffer can be equally accessed from the main SL system and from the user written routine. The user-written routine can be debugged separately in Turbo Pascal environment and then translated as a CHN file and executed from SL. This simplifies the development and debugging very much. The user-written procedure is invoked from the system command line.

The system is designed to receive different data in the form of ASCII string. This way of transfer represents an easy link to large set of other programs written in different languages. These commands are used to transfer data from other programs. For example with SL we use the standard library of signal processing programs for digital filtering, FFT etc. [1].

Conclusion.

The SL system was successfully used in our department to solve various tasks. The SL is a very simple system but gives the user many capabilities and simplifies the overall development.

References.

[1] Digital Signal Processing Committee of IEEE Acoustics, Speech, and Signal Processing Society, "Programs for digital Signal Processing", IEEE Press, 1979.

[2] TURBO Pascal Reference Manual Version 3.0. Borland International Scotts Valley, California, 1985.

ИНТЕРАКТИВНАЯ ЛАБОРАТОРНАЯ СИСТЕМА ДЛЯ АНАЛИЗА И ОБРАБОТКИ РЕЧЕВЫХ СИГНАЛОВ

С.Л. Гончаров, В.Я. Чучупал

Вычислительный центр
Академия наук СССР
Москва, СССР

РЕЗЮМЕ

Доклад содержит описание программного обеспечения интерактивной системы для анализа и цифровой обработки речевых сигналов на малой ЭВМ общего назначения, оснащенной графическим дисплеем. Система представляет собой комплекс программ, в который входят как библиотека процедур, реализующих выполнение стандартных функций по обработке сигналов, так и программы, позволяющие, в диалоговом режиме, с использованием графического изображения речевого сигнала и его параметров на экране дисплея выполнять анализ речевых сигналов, выделять значения их параметров и сохранять эти значения в базе данных, а также выполнять некоторые функции по обработке речевых сигналов, например, производить анализ зашумленных фонограмм с целью улучшения их качества.

ВВЕДЕНИЕ

Исследования в области автоматического распознавания, цифровой обработки сигналов и экспериментальной фонетике зачастую носят трудоемкий и рутинный характер, так как связаны с большим объемом ручной работы, требующейся, например, при подготовке экспериментального материала, а также оценке результатов работы. С начала 80-х годов в Вычислительном центре Академии наук СССР разрабатывается специализированное программное обеспечение для цифровой обработки речевых сигналов на малых ЭВМ. Целью разработки является максимальное облегчение усилий пользователей при программировании и отладке процедур анализа и обработки речевых сигналов, в частности, создание программного обеспечения рабочего места для анализа речевых сигналов. Полученные к настоящему времени в результате проделанной работы программные средства оправдали ожидания разработчиков, как существенно повысив производительность труда специалистов, так и выполнять исследования, ранее практически невозможные.

АРХИТЕКТУРА СИСТЕМЫ

На рис. 1 изображена схематически архитектура системы, включая как программное обеспечение, так и аппаратную часть. Технической базой для рабочего места явился измерительно-вычислительный комплекс, включающий в себя мини ЭВМ, аппаратуру ввода-вывода речевых сигналов и графический дисплей. Всё разработанное программное обеспечение функционирует в среде операционной системы реального времени. Программное обеспечение построено по иерархическому принципу. В этом смысле систему можно рассматривать как совокупность четырех основных компонент - "уровней". Уровни организованы таким образом, что модули верхних уровней ссылаются при работе на модули нижних уровней и могут обмениваться информацией с ними; однако при этом передаваемые параметры стандартизированы так, что структура модулей нижнего уровня остается скрытой от модулей верхнего уровня. Модулям нихних уровней, в свою очередь, недоступна информация о существовании более высоких уровней. Внутри каждого уровня соблюдался модульный принцип построения программ, в соответствии с которым каждая программа выполняет свою, достаточно автономную функцию обработки и только ее. Подобная архитектура обеспечила большую степень независимости программного обеспечения, позволив сравнительно легко приспосабливать его как к новым аппаратурным ресурсам, так и к новым областям применения.

БАЗОВОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

Эта часть программного обеспечения представляет собой совокупность пяти библиотек (объектных модулей), на основе которых создано всё программное обеспечение более высокого уровня. Библиотека стандартных подпрограмм цифровой обработки речевых сигналов содержит стандартных подпрограмм численных преобразований (алгоритмы вычисления БПФ для комплексных, действительных и целочисленных данных, оценки спектральных характеристик сигнала, огибающей спектра, частоты основного тона, параметров линейного предсказания речевого сигнала, синтеза цифровых фильтров и фильтрации речевых сигналов и т.п.). Библиотека подпрограмм матричной арифметики включает набор процедур для выполнения матричных операций на мини ЭВМ, не имеющей спецпроцессора. Использование этих подпрограмм позволяет в 4-6 раз ускорить процесс вычислений по сравнению с обычными процедурами, написанными, например, на языке Фортран-4. Библиотека подпрограмм ввода-вывода речевых сигналов предоставляет средства для обслуживания АЦП и ЦАП. При этом ввод (вывод) данных может производиться как в оперативную память ЭВМ, так и в указанный файл на носителе прямого доступа. Речевой сигнал в последнем случае хранится в формате обычного (двоичного) файла подсистемы управления файлами операционной системы.

Набор подпрограмм управления окнами включает в себя процедуры, предоставляющие наиболее удобный интерфейс между алгоритмами обработки речевых сигналов (использующих обычно пошаговую обработку данных) и самими данными, содержащимися в файлах на носителе прямого доступа. В алгоритмах обработки речевых сигналов оценка значений параметров сигнала и обработка сигнала обычно выполняются в пошаговом режиме, когда речевой сигнал рассматривается как последовательность перекрывающихся между собой сегментов ("окон") данных. Каждый такой сегмент обрабатывается алгоритмом отдельно и, в большинстве случаев, обработка сегментов выполняется однотипным образом. С точки зрения подпрограмм управления окнами файл, содержащий речевой сигнал (или файл, состоящий из параметров речевого сигнала) имеет структуру матрицы, каждая строка которой - вектор значений параметров речевого сигнала, соответствующий определенному моменту времени. В частном случае, когда в файле содержится речевой сигнал, число элементов в каждой строке равно 1, то есть это - одномерная матрица. Прикладная программа должна содержать описание используемых ею данных и правила, по которым будет выполняться разбиение файла данных на сегменты для обработки. С этой целью в прикладной программе специальным запросом создаются буфера - "окна", в которых содержатся данные. Окна данных могут быть созданы для всех входных, выходных и промежуточных потоков данных алгоритма. Поскольку данные описаны как файлы, то последовательность действий по описанию и управлению потоком данных в алгоритме включает в себя следующие этапы:

- описание окон в прикладной программе;
- установление связи между окнами и отображенными в этом окне файлами;
- инициирование операций передачи данных между окном и связанными с ним файлами в соответствии с требованиями алгоритма.

Программа сообщает размер каждого окна (в секундах), шаг, с которым оно смещается по данным, по сигналу (он может изменяться, как и длина окна в процессе обработки), тип окна (окно для чтения и записи данных или только для чтения или записи), а также характеристики отображаемых в этом окне данных: частоту дискретизации, количество параметров и т.п. После того, как программа выдаст запрос на установление связи между созданным ею окном и файлом данных, она может полностью контролировать все функции по обмену данными, с помощью запросов типа "позиционировать окно", "сдвигать окно вверх", "сдвигать окно вниз", "получить время позиционирования окна" и т.п. Когда программа выдает запрос "позиционировать окно", она сообщает точку позиционирования (в секундах от начала записи) и в область окна выполняется передача данных из файла данных, которые соответствуют указанному программой времени. Сдвиг окна на шаг вверх подразумевает смещение данных в буфере таким образом, что начало буфера смещается по файлу данных на промежуток времени, равный шагу смещения окна в сторону возрастания времени. Сдвиг окна вниз аналогичен сдвигу вверх, но при этом окно смещается в сторону уменьшения времени, то есть к началу файла данных. Программа может получить время, которое соответствует началу расположения окна в файле данных выдать запрос "получить время".

При таких условиях работы алгоритм, который реализует пошаговую обработку сигнала от начала до конца, должен просто после очередного цикла обработки сигнала выдавать запрос на сдвиг окна вверх. Описанный выше подход позволил не только резко ускорить процесс создания программ обработки сигналов (включая их отладку), но и обусловил существенную гибкость этих программ за счет отсутствия необходимости в их модификации в случае изменения параметров анализа сигналов (например, длины анализируемого сегмента данных) или характеристик самого сигнала (например, частоты дискретизации). Файлы данных, как уже было упомянуто, могут содержать не только дискретные значения самого речевого сигнала, но также состоять из произвольных параметров речевого сигнала. Этими параметрами могут быть как значения параметров модели речеобразования, измеренные на интервалах анализа, так и более сложные лингвистические характеристики, такие как точки начала или конца определенных фонем, границы участков пауз, смычек, звонких или глухих звуков и т.п. Файлы данных, содержащие значения параметров речевых сигналов, полученных в результате кратковременного спектрального анализа, удобно обрабатывать с помощью представленных выше функций управления окнами данных. Однако для организации работы с данными, характеризующими лингвистические характеристики речевого сигнала, более удобным является дополнительно к этому использование специальных функций системы управления базой данных. Основных функций несколько: это поиск и-го вхождения строки параметров (меток или маркеров) в файл параметров, начиная с указанной точки поиска. В этом случае в прикладную программу передаются времена, соответствующие точкам начала и конца и-го вхождения в файл параметров определенной метки или последовательности меток. Таким путем прикладная программа может, например, последовательно получить доступ ко всем сегментам речевого сигнала, содержащим квазистационарный участок ударной гласной "а" или ко всем участкам, соответствующим определенным слогам и т.п. Пользователь может также специальным вызовом занести значение определенного параметра в указанное место (по времени) файла параметров. Сервисные процедуры позволяют получить перечень значений параметров, содержащихся в указанном файле данных. Следует отметить, что базовое программное обеспечение никак не интерпретирует и не использует семантику параметров, содержащихся в файле данных: это целиком определяет сам пользователь.

В "функциональный уровень" (см. рис. 1) попали процедуры, реализующие законченные алгоритмы обработки речевых сигналов. Состав этого уровня зависит от проблемной ориентации системы. В данном случае наиболее интенсивно использовалось программное обеспечение для коррекции зашумленных сигналов. Однако при всех вариантах использования системы в этот уровень целесообразно включать процедуры, обеспечивающие ввод (вывод) речевого сигнала в файлы на дисках ЭВМ, процедуры, обеспечивающие визуализацию речевого сигнала (типа "видимая речь"), процедуры для генерации сигналов специального вида (тестовые сигналы, белый шум, розовый шум, тоновые сигналы), процедуры оценки первичных параметров речевого сигнала, аддитивного наложения сигналов и т.п. Одной из наиболее интенсивно используемых процедур функционального уровня является интерактивная графическая система для анализа речевых сигналов.

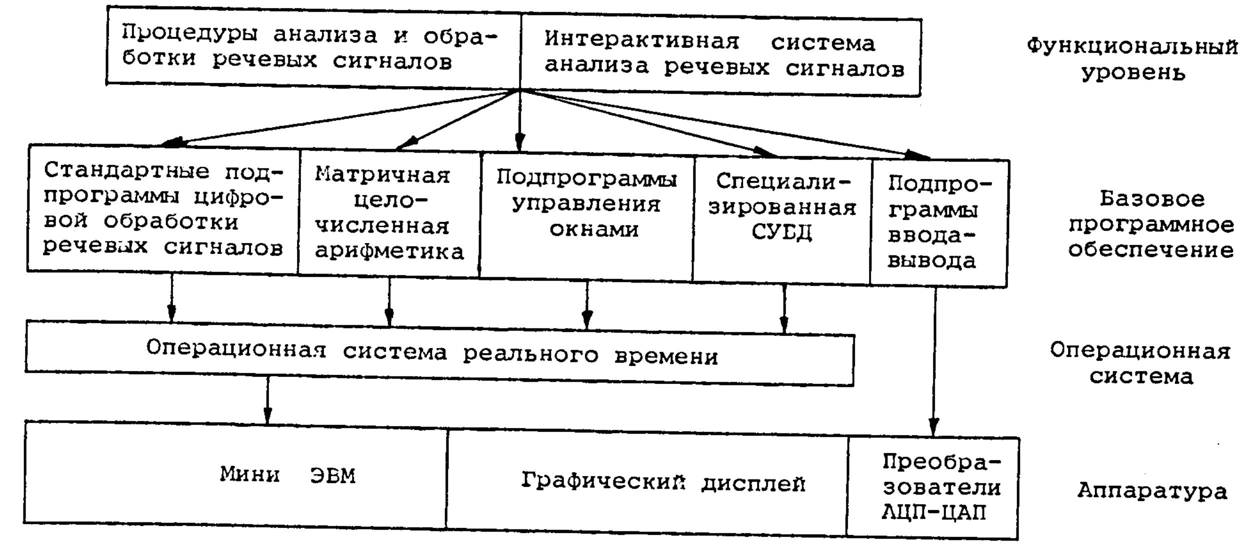


Рис. 1. Иерархическая схема организации диалоговой системы цифровой обработки речевых сигналов.

ИНТЕРАКТИВНАЯ ГРАФИЧЕСКАЯ СИСТЕМА

Интерактивная графическая система предоставляет пользователю возможность выделения, в диалоговом режиме, информативных параметров речевых сигналов, сохранение полученных таким образом значений в базе данных и использование этих значений при выполнении процедур обработки речевых сигналов. Текущая версия системы включает в себя прикладные программы обработки речевых сигналов, которые позволяют выполнять обработку зашумленных речевых сигналов (фильтрацию сигнала из смеси с аддитивным шумом), а также выполнять в диалоговом режиме разметку речевого сигнала, сохраняя в базе данных сведения об основных параметрах сигнала (как-то: значение признака тон/шум, частоты и амплитуды формант, частоты основного тона и т. п.), что может быть полезным при создании специализированных баз данных для тестирования и отладки устройств и алгоритмов обработки и распознавания речи.

Система дает возможность выполнять пользователю следующие действия:

- создавать на экране графического дисплея изображение различных характеристик речевого сигнала: кратковременного амплитудного спектра, временной формы речевого сигнала, сглаженного амплитудного спектра, контура кратковременной интенсивности;
- производить непосредственно на экране графического дисплея указание положения или значений информативных параметров речевых сигналов и запоминать найденные значения в базе данных;
- управлять просмотром сигнала, перемещая смотровые окна в нужном направлении или позиционируя их в интересующей точке;

- изменять тип отображаемых в данный момент характеристик сигнала;
- прослушивать речевой сигнал в указанных пользователем фрагментах;
- указывать границы участка обработки и обрабатывать выделенный участок с помощью заданных пользователем процедур.

Перед началом работы в системе в диалоговом режиме пользователем устанавливаются основные параметры работы с системой, к которым относятся, например, частота квантования, длина и смещение окна, тип весовой функции (прямоугольное окно или окно Хемминга), наличие или отсутствие коррекции верхних частот и т.п.

Анализ сигналов в интерактивном режиме проходит следующим образом. Изображение на экране дисплея можно рассматривать как совокупность нескольких различных представлений сигнала, которые изображаются соответственно, в различных областях экрана дисплея:

- отображение графика кратковременной интенсивности речевого сигнала (окно 3),
- отображение графика речевого сигнала (окно 2),
- отображение амплитудного спектра сигнала (окно 1).

При работе в системе пользователь должен информировать систему о том, с каким окном он работает в данное время, то есть "активизировать" окно. Остальные окна, во время активности одного из окон, пассивны. Существует набор команд работы с окнами. Выполнение команды происходит по нажатию определенной клавиши на функциональной клавиатуре графического дисплея. Система имеет в своем распоряжении следующие команды работы с окном: позиционирование окна, сдвиг окна вверх (вниз) по сигналу, переход к окну 1 (2 и 3).

После того, как окно активизировано, можно переходить к заданию параметров сигнала в интерактивном режиме. Для задания данных в распоряжении пользователя имеется (в каждом окне свой) курсор. Пользователь может указать интересующую его точку следующими способами:

1. Подогнать курсор в нужное место с помощью команд перемещения курсора.
2. Указать элемент изображения световым пером.
3. Сначала "грубо" указать интересующий нас элемент изображения, а затем "тонко" подправить положение курсора.

Кроме команд перемещения курсора (вверх, вниз, влево, вправо) в распоряжении пользователя находятся следующие команды (эти команды выполняются как нажатием клавиши на функциональной клавиатуре графического дисплея, так и указанием световым пером на команду в меню, расположенном на экране дисплея): фиксация точки, привязка к точке, удаление точки.

В описываемую систему встроены две сервисные программы общего назначения, которые требуются практически при всех видах работ. Эти программы выполняют следующие функции: копирование - введенная точка интерпретируется как начало копируемого сигнала, а текущее положение курсора как конец участка копирования; вывод - на ЦАП с целью прослушивания или записи на магнитофон выводится сигнал на указываемом пользователем сегменте. Текущая версия системы включает в себя ряд процедур обработки зашумленных речевых сигналов. Эти процедуры реализованы в виде команд системы, задаваемых как с клавиатуры графического дисплея, так и световым пером в меню на экране дисплея. Система имеет следующие команды:

- "шум" (участок от заданной точки до текущего положения курсора используется для оценки априорных сведений о параметре шума),
- "пауза" (аналогичный участок сигнала интерпретируется как пауза. Соответственно интенсивность сигнала снижается на заданную величину),
- "вычитание" (на аналогичном участке происходит вычитание амплитудных спектров),
- "фильтр" (на аналогичном участке выполняется обработка сигналов квазиоптимальным фильтром для стационарной гауссовой помехи. При этом требуется задание еще одной точки - центра фильтрации).

При работе с речевым сигналом с помощью интерактивной системы обычно используются два типа файлов: это файл, содержащий собственно значение речевого сигнала, которые используются для последующей обработки или анализа речевого сигнала, и файл параметров, в который могут заноситься значения параметров сигнала, найденные в процессе его анализа. К таким значениям могут относиться метки акустических событий (начало и конец участков определенных фонем, пауз, смычек, признаки тон/шум и т.п.). Файл параметров используется при решении задачи накопления в базе данных значений параметров, полученных в ходе ручной или полуавтоматической разметки (анализа) речевых сигналов. В настоящее время решается задача автоматического сбора статистических сведений о накопленной таким образом информации, например, с помощью алгоритмов кластерного анализа.

К системе могут присоединяться новые прикладные программы пользователя. При этом они должны находиться на диске в виде загрузочного модуля. Основная программа системы при этом запоминает свое текущее состояние, поэтому возможен возврат из прикладной программы в систему.

ЗАКЛЮЧЕНИЕ

В докладе описано программное обеспечение интерактивной лабораторной системы для анализа и обработки речевых сигналов. Существующий вариант системы направлен на обработку зашумленных сигналов, однако при некоторых изменениях системы она станет более универсальной. Разработчики видят следующие возможности применения системы:

- полуавтоматическая разметка сигналов. При этом один из сигналов размечается вручную, остальные такие же сигналы (полученные от разных дикторов) размечаются с использованием алгоритмов динамического программирования,
- создание банков фонем, слогов, слов языка. Эти банки данных могут быть использованы в обучающих системах для автоматического распознавания речи.

Возможны также и другие приложения. Таким образом, мы надеемся, что интерактивная система анализа и обработки речевых сигналов является универсальным инструментом в руках исследователей речи.

ЛИТЕРАТУРА

1. В.Я. Чучупал. Диалоговая система цифровой обработки зашумленных речевых сигналов. Диссертация на соискание ученой степени кандидата физико-математических наук. Москва, 1985.
2. В.Р. Женило, Р.С. Иванова, П.В. Миюсов, В.Я. Чучупал. Применение измерительно-вычислительных комплексов для цифровой обработки речевых сигналов. Москва, Вычислительный Центр АН СССР, Сообщения по вычислительной технике, 1985.
3. С.Л. Гончаров, В.Я. Чучупал. Интерактивная лабораторная система для анализа и обработки речевых сигналов. Тезисы докладов и сообщений 14-го Всесоюзного семинара (АРСО-14). 26-28 августа 1986 г. Часть II, с. 70 - 71. Каунас, 1986.

PHONETIC CODING FOR DATA BASES AND EXPERT SYSTEMS

GILBERT PUECH et PIERRE BANCEL

Centre de Recherches Linguistiques et Sémiologiques
Université Lumière-Lyon 2
69500 Bron, France

ABSTRACT

The International Phonetic Alphabet (IPA) is the standard reference as a transcription system. With only minor variants, it is commonly used by linguists to record the pronunciation of languages whether they are supported by an orthographic tradition or not. The scope of this paper is to transpose the IPA to a computer-oriented coding system in order to use phonetic records in data bases and expert systems.

INTRODUCTION

A computer-oriented coding system for the representation of sounds should be viewed as an interface between linguists faced with the representation of a wide range of sounds and a Data Base Management System.

First the code corresponding to each sound must be a key to its major characteristics and, consequently, to the way it patterns with other sounds. The binary distinctive features theory seems to be the natural interface between phonetic analysis and the binary logic of computers. It turns out, however, that there is no clear agreement on how a number of complex or rare sounds should be treated in this approach; furthermore the built-in definition of some features is costly since it precludes some combinations - for instance [+High] is exclusive of [+Low] - or hardly satisfactory to account for some sounds - such as flaps and trills. On the other hand an IPA based classification presents several advantages : it is independant of any particular theory ; it associates phonetic interpretation and a graphic representation in the same table ; it allows a more compact code. This code can be easily converted into a matrix of distinctive features so that the exploitation of the data can be independent of the coding system.

Secondly, the coding system must fit one of the standard formats for computer words. It should also be used to classify phonetically recorded words in the data bases in the same manner as the ASCII code is used to classify orthographically recorded words. If the data base is organized in n-ary trees, the algorithm will find all the relevant information necessary for the equilibration of the trees in the set of codes forming each word.

GENERAL ORGANIZATION

For maximal efficiency, each segment is coded in a short integer (16 bits word) noted by 4 hexadecimal figures. Consonants and vowels are coded independantly of each other, thus it is necessary to know if one given code refers to a consonant or to a vowel before being interpreted. For languages - such as Bantu - in which words are built after a strict syllabic pattern, the data base may determine the fields composing the records as corresponding either to a consonant or to a vowel; in languages where no such syllabic regularity prevails, the first field of the record (a long integer) will in the first byte determine the number of segments included in the record and, in the three following bytes, select the V/C choice (bits 8 up to 31 set to 1 when the segment should be interpreted as a vowel and left at 0 if it is a consonant). Suprasegmental information - stress and pitch - is normally associated with vowels; provision is made however for consonants bearing a tone. A set of diacritics is used to give maximal versatility to this coding system which was designed both for narrow and broad transcriptions. Coding of morpheme boundaries for morphophonemic representations was not examined but could be accomodated.

CONSONANTS

A - Basic consonants are coded in the least significant byte of the short integer. Table 1 yields the phonetic interpretation of the coding and illustrates some of the realizations. The 4 most significant bits correspond to the lines (manner of articulation) and the 4 remaining bits to the columns (place of articulation) :

Phonetic symbol	Code	Phonetic interpretation
b	0041	bilabial voiced stop
m	00C1	bilabial nasal stop
kp	001C	labiovelar unvoiced stop

Sonorants (lines B to F) are assumed to be voiced; implosives and ejectives are respectively voiced and unvoiced. For clicks, which may be voiced, aspirated, murmured etc., further qualification is needed. In order not to have more than 15 places of articulation, some choices had to be made; thus, apico-labial sounds, which are to be found in Umotina [1], are not included in the set of basic consonants but could be handled as a special case (see section F). To facilitate the editing on the lineprinter, it is convenient to have each basic symbol occupy one space only even if it is commonly transcribed as a sequence of two consonants (such as kp or ts).

B - Double consonants, geminates as well as complex segments, are coded in two morae and occupy two spaces:

bb	4141	geminate bilabial voiced stop
mb	C141	bilabial prenasalized voiced stop
nt	C414	alveolar prenasal. unvoiced stop
nts	C474	alveolar prenasalized unvoiced affricate

C - A release, transcribed by a right-adjacent diacritic occupying half a space, is coded in the least significant byte: the most significant bits refer to Table 2; the final hexadecimal zero is a flag indicating that the basic consonant (coded in the first byte) is followed by a release, the interpretation of which is given in Table 2:

kʸ	1B90	velar stop/palatal release
bʷ	41C0	bilabial stop/labiovelar release
dʳ	44F0	alveolar stop/alveolar trill release

Codes which are left free may be defined as necessary.

D - A segment synchronic property, transcribed by a subscribed diacritic, is coded in the most significant byte. The initial hexadecimal zero is a flag indicating that the first byte is to be interpreted as shown in Table 3:

y̰	0CB9	nasalized palatal approximant
z̦	0DA4	lateralized alveolar fricative
m̥	04C1	unvoiced bilabial nasal stop

Provision was made to code the lenis quality on a par with the fortis. However, the lenis quality is assumed to be the unmarked case and it is not associated with a graphic diacritic:

t	0114	lenis t
t̲	0214	fortis t

E - Consonants may be syllabic and bear tones. The syllabicity is coded by the least significant byte set to zero:

m̩	C100	syllabic bilabial nasal stop
s̩	9400	syllabic alveolar unvoiced fricative

Tones on consonants are coded as they are on vowels (see VOWELS, B); tone bearing consonants are assumed to be syllabic.

ḿ	C104	syllabic nasal stop/high tone
m̀	C102	syllabic nasal stop/low tone

F - The overwhelming majority of known consonants may be coded according to the preceeding conventions. However it may be crucial in some languages to handle difficult cases as accurately as possible. We shall resort to the following system: the most significant byte is used as a pointer to a specific filter corresponding to the primary consonant coded in the second byte. One has access, through this filter, to a complementary code, so that the resulting code is extended to 3 bytes; the flag set to detect this situation is the zero corresponding to the least significant bits of the first byte:

ndʳ Prenasalized stop/ trill release	10C4 filter C4/1 : 44F0 extended code : C444F0
ŋ̤ʗ Murmured prenasalized click	10CB filter CB/1 : 0567 extended code : CB0567
ŋʗ Voiced prenasalized click	20CB filter CB/2 : 0467 extended code : CB0467

VOWELS

A - A short vowel - one mora - is coded on a short integer. A long vowel or a diphthong is coded as two morae. The most significant byte corresponds to segmental information. Vowels are plotted on an articulatory space defined by two axes: height (5 degrees) and tongue position in the oral cavity (front, central, back):

	Front	Central	Back
height	1	6	B
	2	7	C
	3	8	D
	4	9	E
	5	A	F

The most significant bits are interpreted as follows:

bit 0 - approximant-like vowel
1 - marked tongue root
2 - nasal
3 - round

The bit 0 is used to mark superclosed vowels (like reconstructed proto-bantu į/ų) or, more generally, the non syllabic part of a diphthong:

əi̯	0A00	8100	diphthong with gliding i
iə̯	0100	8800	diphthong with gliding ə

The bit 1 is used to interpret marked tongue root position (emphatic vowels in the Berber-Arabic domain or the harmonic set of vowels characterized by Advanced Tongue Root in a number of subsaharan languages). Nasality and roundness may combine with this feature:

i	0100	(unrounded) i
u	1B00	(round) ʊ
i̱	2100	nasalized i
u̱	3B00	nasalized u
I̤	4200	ATR I

Basic symbols corresponding to the set of unrounded vowels and of rounded vowels are shown in Tables 4 and 5 respectively.

B - Suprasegmental information is coded in the second byte. Tonal languages use up to 5 levels of pitch, represented henceforth as accents. The code 06 is reserved for a downstepped High:

0101	Falling low	ȉ
0102	Level low	ì
0103	Mid	ī
0104	High	í
0105	Suprahigh	i̋
0106	Downstepped High	ꜜi

Contour tones are coded by reference to their source/target pitch:

0142	Falling high-low	î
0124	Rising low-high	ǐ

The bit 4 is set to 1 if the corresponding tone is floating:

014A	High + Floating low	í̀
012C	Low + Floating high	ì́

Double contours require two morae; we propose the convention that the first mora bear a level tone and the second a contour tone:

0104	0124	Falling-rising long i	íǐ
0102	0142	Rising-falling long i	ìî

C - In order to maximally compact suprasegmental information the bit 0 is reserved for stress:

0180	stressed i	ˈi

If the stressed vowel bears a tone, the code is modified accordingly:

0182	stressed i/low tone	ˈì
01C2	stressed i/falling tone	ˈî

The code A0 is assigned to pitch accent as required by some languages:

01A0	i associated with pitch accent ī

D - Hexadecimal codes 7 and F are left free in our system. Corresponding combinations will be used to account for marked voice quality:

unvoicing	0107	unvoiced i	i̥
	0147	unvoiced i high tone retained	í̥
creaky voice	012F	creaky i/low tone	ḭ̀
breathy voice	0172	breathy i/low tone	ì̤

Special cases may be treated with an extended code as proposed for consonants: a flag (hexadecimal F) indicates that one has to go through a filter table, access to which is given by the code of the vowel mora and a pointer:

01F2 : go to case 2 of the filter table corresponding to vowel i.

Rhotacized vowels, for instance, could be conveniently dealt with in this way.

CONCLUSION

It is indeed possible to rely on the International Phonetic Alphabet to propose a comprehensive and versatile computer oriented coding system. The fact that the code is phonetically motivated makes it particularly attractive for expert systems aiming at comparing data or reconstructing proto-languages.

Reference

[1] P. Ladefoged, "Preliminaries to Linguistic Phonetics", The Univ. of Chicago Press,1971.

	bilabial 1	labiodental 2	dental 3	alveolar 4	labioalveolar 5	retroflex 6	postalveolar 7	prepalatal 8	palatal 9	labiopalatal A	velar B	labiovelar C	uvular D	pharyngeal E	glottal F
unvoiced consonants 1	p			t	pt	ʈ			c		k	kp	q		ʔ
aspirated 2	ph			th							kh				
ejectives 3	p'			t'							k'				
voiced consonants 4	b			d	bd	ɖ			ɟ		g	gb	ɢ		
implosives 5	ɓ			ɗ							ɠ				
clicks 6	ʘ	ʇ	ʖ			!	ʗ								
unvoiced affricates 7		pf		ts			tʃ					kf			
voiced affricates 8		bv		dz			dʒ					gv			
unvoiced fricatives 9	ɸ	f	θ	s		ʂ	ʃ	ɕ	ç		x		χ	ħ	h
voiced fricatives A	β	v	ð	z		ʐ	ʒ	ʑ	j		ɣ		ʁ	ʕ	ɦ
approximants B	ʋ			ɹ					y	ɥ		w			
nasals C	m			n		ɳ			ɲ		ŋ		ɴ		
laterals D				l		ɭ			ʎ		ɫ				
flaps /taps E				ɾ											
trills F	ʙ			r		ɽ							ʀ		

Table 1

Symbol	Example	Code	Phonetic interpretation
˺	t˺	1410	unreleased
ʰ	tsʰ	7420	aspirated release
ˀ	tsˀ	7430	glottal release
ʸ	kʸ	1B90	palatal release
ᶣ	kᶣ	1BA0	labiopalatal release
ⁿ	tⁿ	14B0	nasal release
ʷ	bʷ	41C0	labiovelar release
ˡ	tˡ	14D0	lateral release
ˤ	tˤ	14E0	pharyngeal release
ʳ	dʳ	44F0	trill release

Table 2

Symbol	Example	Code	Phonetic interpretation
	t	0114	lenis
̱	ṯ	0214	fortis
̥	m̥	03C1	unvoicing
̬	s̬	0494	voicing
̤	b̤	0541	murmur
̹	s̹	0A94	rounding
ˠ	tˠ	0B14	velarization
̰	w̰	0CBC	nasalization
ˡ	zˡ	0DA4	lateralization
ˤ	tˤ	0E	pharyngalization
̰	b̰	0F41	laryngalization

Table 3 -

Unrounded vowels

i	ɨ	ɯ
ɪ	ɪ̈	ɯ
e	ə	ɤ
ɛ	ɐ	ʌ
æ	a	ɑ

Table 4

Round vowels

y	ʉ	u
ʏ	ʉ	ʊ
ø	ɵ	o
œ	ɞ	ɔ
ɶ	ɒ	ɒ

Table 5

THE EFFECT ON F_0 OF THE LINGUISTIC USE OF PHONATION TYPE

IAN MADDIESON
Phonetics Laboratory
Linguistics Department
University of California
Los Angeles, CA 90024. USA

SUSAN A. HESS
Phonetics Laboratory
Linguistics Department
University of California
Los Angeles, CA 90024. USA

ABSTRACT

Phoneticians generally expect that laxer adjustments of the vocal cords will produce lower F_0. Hence, languages with phonological contrasts between syllables with tense (somewhat creaky) and lax (somewhat breathy) phonation would be expected to show a difference in pitch between them. We measured F_0 in several minority languages of China with contrasts that have been described as tense vs lax. Our results show that a pitch difference is only sometimes present. The patterns are, in part, explicable in terms of different phonetic realizations and different diachronic sources of the tense/lax contrast, and in terms of its phonological function.

A tendency for different phonatory settings to be associated with pitch differences has been noted by many observers. For example, Laver (1980), in his discussion of laryngeal tension settings, remarks that "there is a strong possibility that in tense voice the pitch range will be higher than in lax voice". Later he comments that "lax voice tends to be accompanied by a low pitch-range". But he goes on to note that there is nothing necessary about the association of laryngeal tension with pitch, commenting that "it is certainly possible to compensate for these tendencies."

Laver is discussing tense and lax laryngeal settings as attributes of individual voice quality. However, a number of languages use tense and lax phonation for linguistic contrast between vowels. This phenomenon is quite common among languages spoken in Southwestern China and adjoining parts of Southeast Asia. We have been conducting studies of the phonation type contrast in several of these languages, and have reported some of our results elsewhere (Maddieson & Ladefoged 1985, Maddieson & Hess 1986). In the present paper we focus on the relation between F_0 and phonatory tension in five of the languages in question. We hypothesized that pitch would correlate with tension, following the tendency noted by Laver, in languages which did not also have tonal contrasts. In languages with tonal contrasts with a high functional load and phonological systems in which phonatory tension is not an aspect of particular tones, we anticipated that the need to maintain the separation of tonal registers would inhibit this tendency. Instead, speakers would draw on the compensatory mechanisms available to counteract it.

Our data consists of measurements of F_0 from 5 languages - Wa, Jingpho, Yi, Lahu, and Lisu. Wa is a non-tonal language of the Mon-Khmer family (Diffloth 1980, Qiu, Li & Nie 1980). The others are Sino-Tibetan languages with tonal systems with a high functional load. Yi (Liangshang dialect, Li & Ma 1983) and Jingpho (of Yunnan, Lu 1984) have similar tone systems, distinguishing high, mid and low-falling tones. In these two languages the phonatory contrast is independent of the tone system, although it is limited to

particular syllabic nuclei in Yi. Lisu is usually analyzed as having a 6-tone system in which tense phonation is characteristic of two of the tones (Mu & Duan 1983). These two tones are mid-level and mid-falling, and can be matched with two of the "lax" tones, also mid-level and mid-falling. Mu and Duan transcribe the pitch height of the tense tones as 44 and 42, and the paired lax tones as 33 and 31, implying that the "tense" tones are indeed higher. Lahu has a system of seven tones, two of which are variously described as being checked by a glottal stop (Matisoff 1973), or having tense vowels (Ma 1984). These two tones, high-falling and low-falling, can be matched with two of the tones that occur non-checked or lax. Whereas Matisoff gives the same pitch values for tense and lax tones, Ma transcribes the tense tones as 54 and 21 but the lax tones as 53 and 31, indicating a smaller pitch range for the tense ones.

3 speakers of each language were recorded with the assistance of Ren Hongmo. The speakers read a wordlist containing 8-10 pairs of monosyllabic words with a minimal tense/lax contrast. Each list was read twice, giving 48-60 examples of each contrast (except for Lahu where only one repetition was recorded, giving 30 cases). F_0 was measured at the onset and offset of the vowel from narrow-band spectrograms. If a more extreme value of F_0 occurred after the vowel onset that value was also measured.

The F_0 measurements in each language were examined in a 3-way analysis of variance, specifying speaker, word pair and tension as main effects. In Table 1 the mean onset and offset F_0 values are shown for the tense and lax vowels in each language. Significant differences (at the .0001 level) are printed bold. All other tense/lax differences are not significant (fall below the .05 level). Measurements of the peak F_0 value did not show a different pattern from those made at the onset, hence these measures are not reported.

Table 1. F_0 measures on tense and lax vowels.

	Wa		Jingpho	
	onset	offset	onset	offset
"tense"	146	112	**157**	128
"lax"	145	115	**145**	126
	Lisu		**Lahu**	
	onset	offset	onset	offset
"tense"	147	119	213	**195**
"lax"	148	122	214	**126**
	Yi			
	onset	offset		
"tense"	**157**	153		
"lax"	**152**	154		

In Wa, words in citation form are spoken with a falling intonation. No pitch difference between tense and lax vowels was observed at either the onset or offset of the vowel. On the other hand, in Jingpho, a significant pitch difference at the vowel onset was observed. The Jingpho wordlist includes pairs of words with all three tones, but pairs with low-falling tone predominate (6 out of 10). Because of this, the mean offset value is low. The word pairs examined in Yi were all mid-level tone, hence onset and offset values are close. The onset F_0 differs between tense and lax syllables by a small but highly significant amount in Yi. In Lisu there is no significant difference at either onset or offset, despite Mu & Duan's indication to the contrary. Since phonatory tension is a property of particular tones in this language we had expected no effort to avoid a pitch distinction. Lahu shows a significant difference in F_0 at the vowel offset. The mean offset value in the two lax falling tones is considerably lower than in the tense tones.

Our results are thus generally counter to our hypothesis, which predicted that an F_0

difference would occur in the nontonal language Wa, and in Lisu and Lahu where phonation type is an aspect of tone, but not in Jingpho or Yi where phonation type is independent of tone.

Are there historical or synchronic facts about these particular languages which enable us to understand this result? Jingpho and Wa share a similar historical origin for the tense/lax contrast, namely, the somewhat breathy lax syllables are those which used to have initial voiced consonants. However, synchronically, the phonation type contrast is more salient in Jingpho than it is in Wa. We have used the difference in amplitude between the second harmonic and the fundamental, $H_2 - F_0$, as our measure of phonation type. This measure has a higher value for tenser phonation than for laxer phonation (Maddieson & Ladefoged 1985). In Wa the mean difference in the $H_2 - F_0$ measure between lax and tense vowels is just under 2 dB, whereas in Jingpho it is just over 7 dB. In addition, the tense/lax contrast in Wa is accompanied by some vowel quality difference: tense vowels have a higher first formant than lax ones, i.e. they are lower in the perceptual vowel space. In Jingpho, vowels in tense and lax syllables do not differ. It may therefore be the case that in Wa the small pitch difference that might have been expected from the not-very-salient phonation type contrast is counteracted by the effect of vowel lowering in tense syllables. In Jingpho on the other hand, the phonation type contrast is made salient enough so that the conditioning environment for any allotonic variation can be readily recognized.

Lisu developed tense phonation in syllables which were originally checked (i.e. stop-final). In Lisu we found that the mean difference in the H_2-F_0 measure between tense and lax was about 3 dB, confirming the existence of a moderately salient phonation type difference. Since there is no pitch difference, this suggests that the system should be reinterpreted as one with four tones in which a phonation type contrast operates within two of the tones, rather than as one with six tones, two of which have a marked phonation type.

Lahu shows no reliable evidence of a phonation type difference based on the measure we have used, nor is there usually any auditory impression of one. Instead, in the historically checked syllables, a final glottal stop usually occurs and the vowel is considerably abbreviated (about 275ms shorter than in "lax" syllables). The much lower offset pitch in the two falling "lax" tones seems simply due to their much greater length; the pitch continues to fall and reaches a much lower level. In Lahu, phonation type is only marginally involved in syllabic contrasts. Duration, extent of pitch change and glottal stop are more central to the contrast which has been described as "tense" vs "lax". Matisoff's representation of the "tense" syllables as having a final glottal stop is more accurate than Ma's account, though Ma correctly indicates the greater pitch range of the "lax" (unchecked) syllables (cf Hombert 1983).

Yi is again somewhat different. Although the difference between "tense" and "lax" vowel pairs is quite distinctive, with an auditorily "harsher" quality for the tense members, the $H_2 - F_0$ measure does not distinguish them. Perhaps this measurement is simply not appropriate for detecting phonatory differences in the rather unusual range of "fricative vowel" segments found in Yi. We think that it is more likely that the tense/lax contrast is produced in a different way here. We speculate that the "tense" vowels employ a supralaryngeal mechanism like that used in the "strident" vowels found in some of the Khoisan languages, which involves a narrowing between the base of the epiglottis and the upper part of the arytenoid cartilages. The use of this mechanism in !Xoo has been described in some detail by Traill (1985). Traill has listened to our Yi recordings and agrees that

there is an auditory similarity between the strident vowels of !Xõõ and the tense vowels of Yi. However, in !Xõõ, strident vowels have somewhat lowered pitch, rather than the slightly higher pitch found in Yi "tense" vowels.

In the meantime, we find that, particularly in the data from Jingpho, we have provided a phonetic basis for a different hypothesis. This is the diachronic hypothesis that tonogenesis and splitting of tones in tone languages can arise from phonation type contrasts on vowels, as has been proposed by Pulleyblank (1978, 1984) for Chinese. Previous work has concentrated on consonantal sources for tones, and the effect of contrasts on vowels has largely been ignored. We now see that such effects can be significant. However, as data from Wa and Lisu demonstrate, phonation type may be contrastive in vowels without any accompanying pitch differences.

References

Diffloth, Gerard. 1980. The Wa Languages (Linguistics of the Tibeto-Burman Area 5.2). California State University, Fresno.

Hombert, Jean-Marie. 1983. A brief encounter with Lahu tones. Linguistics of the Tibeto-Burman Area 7.2: 109-111.

Laver, John. 1980. The Phonetic Description of Voice Quality. Cambridge University Press, Cambridge.

Li Min and Ma Ming. 1983. Liangshan Yiyu yuyin gailun [Description of the sounds of the Liangshang Yi language]. Sichuan Minzu Chubanshe, Chengdu.

Liu Lu. 1964. Jingpoyu gaikuang [Brief description of the Jingpho language]. Zhongguo Yuwen 132: 408-4177.

Ma Shice. 1984. Lahuyu gaikuang [Brief description of the Lahu language]. Minzu Yuwen 1984.3: 70-80.

Maddieson, Ian and Peter Ladefoged. 1985. "Tense" and "lax" in four minority languages of China. Journal of Phonetics 13: 433-454.

Maddieson, Ian and Susan Hess. 1986. "Tense" and "lax" revisited: more on phonation type and pitch in minority languages of China. UCLA Working Papers in Phonetics 63: 103-109.

Matisoff, James. A. 1973. A Grammar of Lahu. University of California Press, Berkeley & Los Angeles.

Mu Yuzang & Duan Ling. 1983. Lisuyu gaikuang. [Brief description of Lisu language]. Minzu Yuwen 1983.4: 72-80.

Pulleyblank, E. G. 1978. The nature of the Middle Chinese tones and their development in Early Mandarin. Journal of Chinese Linguistics 6: 173-203.

Pulleyblank, E. G. 1984. Middle Chinese: a Study in Historical Phonology. University of British Columbia Press, Vancouver.

Traill, Anthony. 1985. Phonetic and Phonological Studies of !Xõõ Bushman (Quellen zur Khoisan-Forschung 1). Helmut Buske, Hamburg.

Qiu Efeng, Li Daoyong & Nie Xizhen. 1980. Wayu gaikuang [Brief description of the Wa language]. Minzu Yuwen 1980.1: 58-69.

THE CORRELATION OF THE TENSE-LAX CONSONANTS IN SOME RUSSIAN DIALECTS AND IN OTHER SLAVIC LANGUAGES

LEONID KASATKIN

The Lenin State
Pedagogical Institute,
Moscow, USSR, 119435

ROSALIA KASATKINA

The Russian Language Institute
of the Academy of Sciences of
the USSR, Moscow, 121019

ABSTRACT

The speech material of the Northern Russian dialects was investigated. A complex of the phonetic phenomena was found testifying to the existence of the correlation of the tense-lax consonants in those dialects. Since analogous phonetic features are observed in the Western and Southern Slavic languages, it may be suggested that the peculiarity discovered in the Northern Russian dialects is pra-Slavic and pra-Indoeuropean heritage.

1. As we know, the consonants of Standard Russian are opposed on the basis of voicelessness-voiceness. The voiced consonants differ from the voiceless ones also by the level of tenseness: the voiceless consonants are more tense. This is manifested in the greater tension of the muscles of the articulatory organs.
One of the most difficult tasks of instrumental phonetics is the direct establishment of the level of tenseness. However, one can judge of the degree of tenseness or nontenseness on the basis of some inderect data. Specifically, the tense consonants compared to the lax ones are longer and the noise constituting them is more energetic /1/.
In Standard Russian the feature of tenseness is closely related to the feature of voicelessness, and the feature of nontenseness - to the feature of voiceness: tense consonants are voiceless and lax ones are voiced. It should also be borne in mind that voiceness-voicelessness is the major feature in the opposition, while tenseness-nontenseness is an accompanying feature /2/. In some languages these phonetic features are correlated otherwise than in Russian. In such languages tenseness-nontenseness lies at the bottom of the opposition. As examples one can cite English, French, German, Finnish, Estonian and many other languages /1, 3/.
2. So far no Russian dialects have been described where the principle of tenseness-nontenseness of consonants manifested itself differently than in Standard Russian. Such dialects are to be found on the river of Mezen in the Leshukon district of the Arkhangelsk region. Our primary auditory impression was checked instrumentally when the length of consonants of these dialects was measured.
2.1.1. According to the data recieved by Zlatoustova for Standard Russian the length of voiceless fricative consonants in the intervocalic position can vary within the range of 167 mc to 213 mc. The voiced consonants show a variation from 93 mc to 127 mc. The ratio of the length of voiced consonants to that of the voiceless ones is approximately 0.56-0.65 /4, p. 57/.
The proportion of voiceless and voiced fricative consonants in the intervocalic position in the Mezen dialects differs from that in the literary language. The difference is a greater contrast in their length. Thus the length of intervocalic [ʃ] varies within the range of 95-100 mc; while [ʒ] in the same position is characterised by the length of 45-59 mc; the temporaly characteristics of the intervocalic [s] are from 110 to 180 mc, and those of [z] - from 50 to 60 mc. The ratio of the length of voiced consonants to that of the voiceless ones in the Mezen dialects is about 0.46 on the average.
2.1.2.An even greater difference between the literary language and the Mezen dialects can be observed in the stops which are longer in the Mezen dialects.
The length of the voiceless stops differs from language to language. In some languages these consonants have a longer phase of contact, which results in geminates. Estonian and Finnish are examples of this phenomenon. In other languages (English, German) the occlusive consonants have a longer postexplosive phase leading to aspirated consonants.
Both types of prolonging of the voiceless stops can be observed in the Mezen dialects. Thus sometimes these consonants are pronounced with a long contact: [a p̄'ímof, 'ētogo, potō'lok, ka'k̄oi]. But more often the length of the voiceless stops [p, t, k] and [p', t', k'] appears in aspiration: ['phom'or, na'phal, 'naphol; tham, 'thoʒə, photho lok; a'khak, khu'dy, a p'hec-tu, ku'p'hila, 'n'ep'hili; 't'hanut, 'mat'hi; muʒy'k'hi].
In Standard Russian the duration of postexplosive phase of the voiceless stops is quite insignificant: [p, t] - 20 mc, and [k] - 35-40 mc /5/. If one takes into consideration the fact that the duration of [p, t, k] in the intervocalic position varies from 153 to 200 mc /4, p. 571/, then the postexplosive phase of [p, t] is equal to 0.1 of the length of the whole consonant and that of [k] - to 0.17-0.25. According to our data the duration of the postexplosive phase of [p] in the Mezen dialects is 42-95 mc, that of [t] is 65-70 mc, and that of [k] - 54-76 mc. The measurement of their relative length showed that the postexplosive phase of these consonants may constitute from 0.4 to 0.7 of the entire length of the consonant.
2.2.1. One can also see the difference between Standard Russian and the Mezen dialects in the proportion of the consonant length in clusters.
In Standard Russian the first consonant of the cluster is typically shorter than the second one /4, p. 59/. This regularity is proved by our measurements of the consonant length in such groups as [ks], [sk], [ps], [sp], [ksh], [shk], [kt], [ft], [gz], [zg], [zb'], [db]. There is a law in Standard Russian according to which the first consonant cannot be longer then the second one even if the first consonant represents the combination of two identical phonemes: the long consonant loses its length when it occurs beside another consonant; compare: классы [s:] - классный [s] /6, p. 136/.
The situation is quite different with the Mezen dialects, where the first consonant may be much longer than the one which follows. Compare: [uʃ'la, uʃ kom, 'tʃ' is̄ ta, l'es̄'na, p'es̄'kom, fsu'botu]. The length of the voiceless stops in clusters as well as in the intervocalic position may come up in aspiration or in the longer contact phase: [okh'no ; nak'laz'da]. The first consonant is longer than the second one even in the case when the former is a sonorant and the latter a voiceless consonant, while in the intervocalic position the voiced consonants including sonorants are much shorter than the voiceless ones. The average length of the second consonant compared to the first one varies from 0.4 to 0.7. The voiceless stops are nonaspirated in the postconsonant position. Therefore the first or the second position of the consonant in the clusters differs to the tenseness-nontenseness. From this point of view the position of the first consonant in the group is strong, and the second one is weak.
2.2.2. Another peculiarity of the Mezen dialects that is the progressive devocalization of the sonorants. This phenomenon is observed both in the middle of the word and in juncture: ['proʃlyi, 'utrom, 'pr̥'amo, d'ek'r'et, sus'l̥on, ûʃ'l̥'i, 'vyp'jut, soʃ'jôʃ]. The sâme devocalization is observed in the sounds of [v, v'], which are pronounced in the dialects; this is also true of the more ancient [w, w']: [kf̂am, dakf l'e'sax, vo'z'it'f'lotkax, sfo'jo].
Completely voiceless sonorants and [f, f'] according to [v, v'] occur much rarer than partially devocalized sonorants. The instrumental analysis of these sounds showed that such sonorants have voiceless beginning and voiced ending. The degree of the devocalization of the sonorant and [v] in the position after the voiceless consonant in the Mezen dialects depends on the force of tension of the speech organs. When used emphatically or in the strong phrase positions the sonorants are devocalized for the greater part of their duration and the sound [f] is pronounced instead of [v]. In other cases the devocalization may extend over the initial phase of the second consonant only. There may be no progressive devocalization of these sounds in the weak phrase positions.
The strong voiceless consonants may influence not only the next sonorants but also the vowels. In such cases vowels are pronounced without voice though preserving the rest of their typical characteristics: [pr̥ʒy'la, pr'iʃlo, 'wyp'ito]. This effect can be observed frequently at the end of syntagma. Sometimes several successive words may be pronounced as if they were whispered, with the strong tension and intensive noise.
In the group of two consonants, as it has been shown above, the first consonant is tense and the second one is lax. That is why if the first sonorant or [v] is following the voiceless consonant the progressive devocalization is observed quite frequently. It almost never happens if the sonorant or [v] is placed after two voiceless consonants: the second sound is lax, it cannot assimilate the next sonorant and [v]; compare: ['trojo-'stroim].
2.3. The prolongêd consonants in the Mezen dialects frequently occur at the end of a word before a pause: [l'es̄, bo'jus', mox̄, 'v'id'iʃ̄]. The stops are pronounced with a long contact and explosion: [i d'ot̄, pesok̄, or with aspiration: ['thuth, o'p'et'h, poto-'lokh]. Quite frequently the voiceless stops are implosive. This may evidently be explained by the fact that the general ab-

atement of the intensity at the end of a syntagma weakens the end of the consonant as well that is why the strength of the contact is greater than the strength of the explosion and the explosion does not take place.
The voiced consonants in the Mezen dialects are lax. They are much shorter than their voiceless correlates. Besides nontenseness manifests itself in the common flabbiness of their articulation. We have often noted the pronunciation of [j] and [ð'] instead of [d]: ['lajila, buðot]; [ð] in place of [d]: ['eðak]; [w, w'] in place of [b, b']: [nara'wotu, w'un'tom]; [ɣ] instead of [g]: ['mnoho].
2.4. In some cases in the Mezen dialects the pronunciation of voiced consonant in place of the voiceless ones and vice versa can be observed and also the pronunciation of semivoiced consonants in place of voiceless and voiced ones: ['star'in ga, (g < k), po'tumat' (t < d), po'rato, za'gadəvəl].
2.5. Implosive consonants, spirantization of voiced explosive consonants, interchange of voiced consonants and voiceless ones and the existence of semivoiced consonants have been noticed in different Northern Russian dialects by other dialectologists. The auditioning of the tapes of the Northern Russian dialects accumulated in the Laboratory of experimental phonetics of the Russian Language Institute of the USSR Academy of Sciences showed that they share some other features with the Mezen dialects which have been described above.
3. All this testifies to the fact that in the Northern Russian dialects there exists opposition on tenseness-nontenseness, but not on voiceness-nonvoiceness as is the case in other Russian dialects and the literary language.
When making phonological conclusions some phoneticians proceed from the principle of phoneme neutralization /3,7/. However the fact of neztralization as such cannot always clarify the nature of the phonetic opposition. Thus [t] and [d] coincide in the sound [t] in the final position both in Russian and in German. Yet in Standard Russian the opposition on voiceness-nonvoiceness is considered to be neutralized in the final position, while in German the opposition of tenseness-nontenseness is neutralized in a tense variant. The fact of neutralization is an evidence that the phonemes are paired and that they are opposed on one distinctive feature. But it may mean nothing as to the nature of this feature. The Mezen dialects as well as the majority of the Northern Russian dialects do not differ from other Russian dialects from the point of view of the nature of neutralization of the consonants discussed above. Here the noise consonants cannot be distinguished in the final and preconsonant position. At the end of the word and before the voiceless consonants they turn into voiceless consonants and before the voiced consonants they turn into voiced ones.
The difference between the two types of dialects lies in how the contrast of the opposed phonemes in the absolutely strong position is realized. In some dialects as well as in Standard Russian the contrast of the consonants on voiceness-nonvoiceness is more evident than on tenseness-nontenseness. In other dialects the contrast of the consonants on tenseness-nontenseness is more evident than on voiceness-nonvoiceness. That is why the opposition of these consonants is rooted in tenseness-nontenseness. The feature which forms the basis of consonant opposition in the absolutely strong position may give up it's place to some accompanying principle under other conditions. Thus in Standard Russian the difference between [p] and [b], [t] and [d], [s] and [z], etc. in whispering, when there is no voice, is evident only from tenseness or nontenseness of the corresponding sounds /6/. In those Northern Russian dialects where the leading principle of phoneme opposition is usually tenseness-nontenseness, in the postconsonant position, where voiceless stops lose aspiration and fricative consonants lose their length, the major contrast between the corresponding sounds is on voiceness-nonvoiceness.
4. What is the origin of the dialect peculiarity described above? Speaking about the vocalization of the voiceless consonants in the intervocalic positions and the existence of the semivoiced consonants some investigators proposed that it is a feature of the Finnish substratum /8/. This proposal has some validity. It is possible that the other features of the described complex are also of Finnish origin.
However there is some counter evidence too. The Komi Republic Academy of Sciences gave us an opportunity to listen to the tapes of different Komi dialects including the dialects on the river of Mezen, neighbouring on the Russian Mezen dialects. In none of these tapes could we find the most typical feature of the Russian Mezen dialects - aspiration of the voiceless stops. Yet some of the manifestations of the opposition on tenseness-nontenseness in the Komi dialects do exist, for example the prolonging of the first consonants in clusters.
There may be another explanation of the described Northern Russian phenomenon. Many indoeuropean languages have the same features. Thus for example the tense voiceless consonants significantly exceed in their length the lax voiced ones; the aspiration of the voiceless stops occurs at the beginning of the word and in the intervocalic positions (while it is absent in the postconsonant position); cf. also the progressive devocalization of the sonorants, the prolonging of the ending consonants, the spirantization of the voiced stops in English and German /9/.
Many of the described phenomena are known in the Slavic languages. According to our data [p, t, k] in Polish are more tense than in Russian. The voiceless stops are aspirated in Polish. There is also the progressive devocalization of the sonorants in some Western and Southern Slavic languages /10/. For Czech the relevance of the opposition on the "lenes-fortes" of the consonants was discussed /11/.
Consequently the discussed features of the Northern Russian dialects connected with the opposition of the consonants on tenseness-nontenseness, may be one more feature linking the Northern Russian dialects with the Slavic West. This feature may be praindoeuropean.

References

/1/ Р.Якобсон, Г.Фант, М.Халле. Введение в анализ речи. Различительные признаки и их корреляты. Гл. П. Опыт описания различительных признаков. - В кн.: Новое в лингвистике. Вып. П. М., 1962.
/2/ Л.Г.Зубкова. Фонетическая реализация консонантных противоположений в русском языке. М., 1974, с. 10-11.
/3/ Н.С.Трубецкой. Основы фонологии. М., 1960.
/4/ Л.В.Златоустова. Фонетическая структура слова в потоке речи. Казань, 1962.
/5/ Л.В.Бондарко. Звуковой строй современного русского языка. М., 1977, с. 143.
/6/ М.В.Панов. Современный русский язык. Фонетика. М., 1979.
/7/ П.С.Кузнецов. О дифференциальных признаках фонем. - В кн.: А.А.Реформатский. Из истории отечественной фонологии. М., 1970, с. 491.
/8/ В.В.Колесов. Севернорусские чередования согласных, парных по глухости-звонкости. - Вестник ЛГУ, 1963, №2. Серия истории, языка и литературы, вып. 1, с. 108; Р.Ф.Пауфошима. Некоторые вопросы, связынные с категорией глухости-звонкости в говорах русского языка. - В кн.: Экспериментально-фонетическое изучение русских говоров. М., 1969, с. 214.
/9/ A.C.Gimson. An introduction to the pronunciation of English. London, 1970, p. 159, 161, 164; В.М.Жирмунский. Немецкая диалектология. М.-Л., 1956, с. 251-252,254.
/10/ B.Hála. Uvedení do fonetiky češtiny na obecné fonetickém zakladě. Praha, 1962, s. 362; J.Belič. Nástin české dialektologie. Praha, 1972, s. 59; А.М.Селищев. Славянское языкознание, т.I. Западнославянские языки. М., 1941, с. 300,327, 335.
/11/ Y.Vachek. K znělostnímu protikladu souhlásek v češtině a v angličtině. - Studie ze slovanské jazykovědy. Praha, 1958.

ON THE PHARYNGEALIZATION IN TUNGUS-MANCHU LANGUAGES

GALINA RADCHENKO

Novosibirsk, USSR 630090

ABSTRACT

The paper presents some results of the experimental study of the Nanay and Udehe phonetic systems. The obtained results concern the nature and the function of the pharyngeal/laryngeal phonemes and their influence on the vowel and consonant patterns. The proposed approach allows a simple solution of some disputable phonomorphological phenomena in Tungus-Manchu languages. The new interpretation of the vowel patterns of Nanay and Udehe languages is given. It is shown that pharyngeal/laryngeal phonemes are marked by functional ambiguity, serve as means of distinctive and delimitative function on segmental, supersegmental and phonomorphological levels. The presence of morphonological constructive elements in the phonemic patterns of Tungus-Manchu languages is a feature typical of the syllabomorphemic language type.

INTRODUCTION

Pharyngealization has wide phonetic manifestation ranging from aspiration of consonants to pharyngealized accent in Tungus-Manchu languages. The pharyngeal / h / was found in some Tungus-Manchu languages: Evenki, Even, Solon and some dialects of Oroch /1/. The occurences of aspiration of voiceless stops were mentioned in certain dialects of Evenki and Even /2/. The aspirated and glottalized vowels with the glottal stop or expiration of breath in the middle of vowel phonation were described by E.R. Shnieder /3/. But on the whole, references to the pharyngealization in Tungus Manchu languages are scanty.
The present paper concerns the phonetic nature, phonological status and function of the pharyngealization in Nanay and Udehe languages, which enter the southern group of Tungus-Manchu languages. The investigation is based on the data obtained experimentally. The list of 300 Nanay and Udehe words was read by 10 informants, about 45-50 years old not phonetically trained, and all of them unaware of the purpose of the experiment. Recordings of this material were made by means of oscillograph, intonograph, at 250 mm/sec and were also treated by spectrograph.

ACOUSTIC MANIFESTATION

It is traditionally accepted that there are voiced and voiceless stops in Tungus Manchu languages. On the acoustic spectra of the Nanay and Udehe words the two sets / bdg / and / ptk / initially are produced with silent closure intervals and ought to be classified as voiceless whereas in medial position / bdg / are voiced and / ptk / are voiceless. The consonant spectra of / ptk / are characterized by postaspiration which manifests itself as higher frequency noise /fig.1/ According to L.Lisker and A.S.Abramson the difference between voiced and voiceless consonants is in the timing of voice onset relative to release: / bdg / are distinctevely marked by low frequency harmonics preceding the burst of the release and / ptk / are distinctevely marked by an interval of higher frequency noise immediately following the burst/4/. For Tungus-Manchu languages this difference works only in part. In initial position / ptk / and / bdg / are voiceless and / ptk / distinguishes from / bdg / by an audible explosion and an interval of mid-higher frequency noise within the range corresponding to the frequency harmonics of the following vowel, i.e. aspiration /fig.2/.
The pharyngeal / h / and the glottal / ? / occur at the beginning of a stem-morpheme before a following vowel or at the morpheme boundary serving as a word boundary marker, e.g. Nanay: / ?ʌniʌ / 'mother', / ?oɧoka / 'small fish'; Udehe: / ?asa / 'bay', / ?unah / 'fingers', / na?u / 'cock and hen', / inahji / 'dog' On the spectra the glottal stop manifests itself in the delay of F_o from the first and second formant frequencies /fig.3/. This is due to the delay in vioce onset. In the weak position, i.e. between vowels

and as a finale of a syllable the pharyngeal / h / and laryngeal / ? / are realized as sonants: ?~h~w~ŋ~ɣ~j (e.g. Nanay:/wonemi ~ŋonemi~?ōnemi/'long'; Evenki:/ala~alaɣ~alak~alah/'motley'). These vocalized laryngeals are often omitted giving rise to long vowels and diphthongs: Negidal: /adaxu~adaku/ 'twin', Oroch: /adawu~adau/, Udehe: /ad?au/, Ulchi: /adau~adū/, Orok: /adaw~adau/, Nanay: /adao/. In Udehe language we observe the process of transition of / h / and / ? / from phonemic to prosodic level: the first or the last syllables of a stem are marked by the pharyngealized/laryngealized accents. These accents are characterized by the double peak fundamental tone (circumflex) which marks the consonant and the vowel of a syllable /fig. 4-8/. The experimental data have shown that the pharyngealization/laryngealization in Udehe language is the distinctive feature of a syllable but not of a vowel as it was stated on the basis of the auditory analysis by E.R. Shneider.

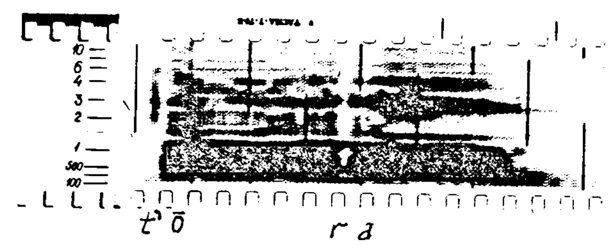


Fig. 1. Spectrogram of the Nanay word /tora/ 'he goes hunting'

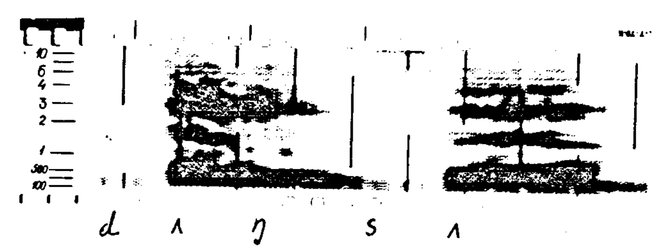


Fig. 2. Spectrogram of the Nanay word /dʌŋsʌ/ 'balance'

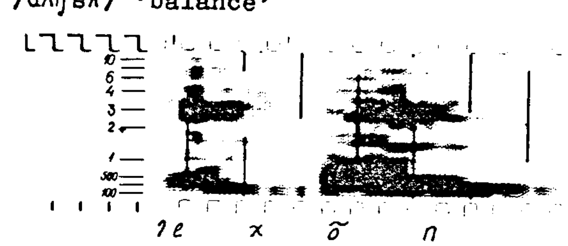


Fig. 3. Spectrogram of the Nanay word /exon/ 'settlement'

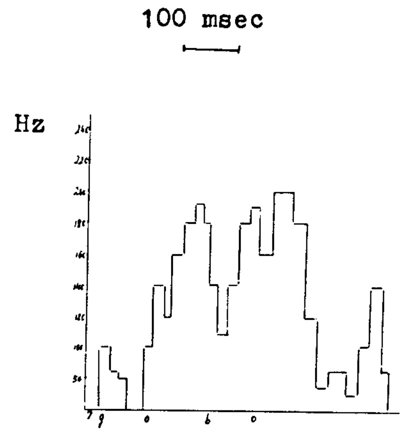


Fig. 4. F_o-curve of the Udehe word /gobo/ 'fly'

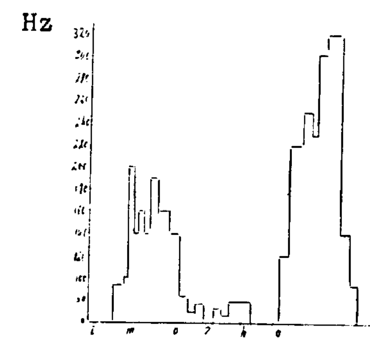


Fig. 5. F_o-curve of the Udehe word /imoho/ 'fat'

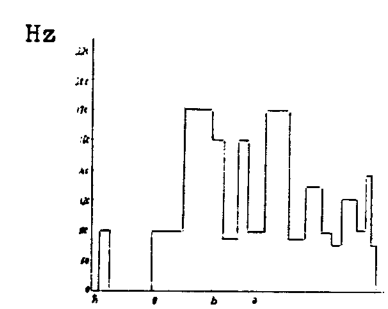


Fig. 6. F_o-curve of the Udehe word /ɦobo/ 'hard'

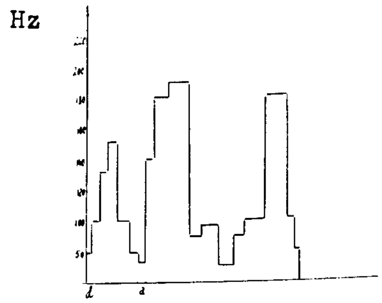


Fig. 7. F_o-curve of the Udehe word /da/ 'cotton wool'

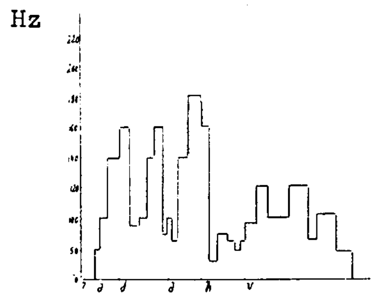


Fig. 8. F_o-curve of the Udehe word /adahu/ 'twin'

FUNCTION

The pharyngeal and laryngeal phonemes function as boundary markers between two morphemes, one ending with a vowel and the other starting with a vowel. These phonemes may be considered as morphonological constructive elements, serving to link a stem and a suffix. For example, the initial glide / w / of many verbal/noun suffixes in Nanay may be omitted. Its presence depends on the syllable structure of a stem. If the syllable of a stem has a long vowel or a diphthong which are always marked by the double peak accent, the morpheme and syllable metanalysis is not possible. If the syllable of a stem is not stressed the morpheme metanalysis is possible, e.g. Nanay: /būwuri/ 'to give', /xolaori/ 'to read'. In Udehe the unstressed syllable of a stem form a fusion with the vowel of the following suffix. In this case the pharyngeal / h / and the laryngeal / ? / which are the markers of the Past Indefinite and the Past Perfect correspondingly, are manifested as pharyngealized/laryngealized accents linking the stem with the suffix: Udehe: /wā?tы / 'they had killed', / ʒaw?a/ 'he had taken'.

In conclusion it should be stated that / h / and / ? / became isolated in the pattern of consonant phonemes. This isolation was due to the functional ambiguity as these phonemes serve both as means of distinctive and delimitative function. The presence of such morphophonemic elements in the language is a feature typical of the syllabomorphemic language type.

REFERENCES

/1/ В.И. Цинциус, "Сравнительная фонетика тунгусо-маньчжурских языков", Ленинград, 1949.

/2/ А.А. Горцевский, "Фонетические трудности при обучении эвенков /тунгусов/ русскому языку", Ленинград, 1939; К.А. Новикова, "Проект единой фонетической транскрипции для тунгусо-маньчжурских языков", Москва-Ленинград, 1961.

/3/ Е.Р. Шнейдер, "Краткий удэйско-русский словарь", Москва-Ленинград, 1936.

/4/ L. Lisker, A.S. Abramson, "Stop Categorization and Voice Onset Time", The Fifth International Congress of Phonetic Sciences. Proceedings, Basel-New York, 1965, pp.389-391.

DAS KONSONANTENSYSTEM DER DOLGANISCHEN SPRACHE (NACH EXPERIMENTALEN ANGABEN)

NATALI BELTJUKOWA

Die Tomsker Staatliche
Kujbyschew-Universität
Tomsk, UdSSR 634010

Die vorliegende Arbeit ist der Erforschung des Konsonantensystems der Sprache der Dolganen gewidmet und sieht die Aussonderung des Konsonantenbestandes, die Bestimmung des Untersystems der Konsonantenphoneme des phonologischen Systems der dolganischen Sprache, sowie die Hauptmerkmale der artikulatorisch-akustischen Grundlage dieser Sprache auf dem Gebiete der Konsonanten vor.

Die Dolganen sind eine kleine (4877 Mann stark) türksprachige Völkerschaft im Hohen Norden, die im Taimyrer (Dolgan-Nenezker) autonomen Bezirk des Krasnojarssker Regions in den Bezirken Dudinka und Chatanga leben.

Die Sprache der Dolganen, die man am Anfang ihres Bestehens als ein Dialekt der jakutischen Sprache mit bedeutenden Archaismen betrachtete, löste sich seit langem unter den Bedingungen der langwierigen historischen Entwicklung vom ganzen System der jakutischen Sprache ab und verlor die Eigenschaften eines Dialekts. Der moderne Stand der dolganischen Sprache läßt uns sie als eine selbständige Sprache betrachten /1/. In einigen Schriften aber wird Dolganisch bis jetzt noch als ein Dialekt der jakutischen Sprache erläutert /2/.

Das Sprachmaterial, welches als Grundlage für die erste Analyse und Verallgemeinerung diente, wurde den Texten (Erzählungen, Märchen), einzelnen Sätzen und Wörtern entnommen, die die Sprachexpeditionsteilnehmer S.Demjanenko, T. Koschewerowa und der Autor dieses Artikels unter der Leitung des Professors A.Dulson 1970 und 1971 festgelegt haben.

Seit 1973 begann der Autor das Sprachmaterial nach einem speziellen Programm zu sammeln, welches im Laboratorium der experimental-phonetischen Forschungen des Instituts für Geschichte, Philologie und Philosophie der Sibirischen Abteilung der Akademie der Wissenschaften der UdSSR unter der Leitung von W.M.Nadeljaew zusammengestellt wurde, sowie auf der Halbinsel Taimyr (während der sprachlichen Dienstreisen des Autors) als auch in Nowossibirsk im genannten Laboratorium im Laufe der Arbeit mit fünf vom Taimyr angekommenen Dolganen.

Die erste Forschung der dolganischen Sprache hat E.I. Ubrjatowa durchgeführt. Sie nennt 21 typische Konsonanten /3/.

Im Lautsystem der nahverwandten jakutischen Sprache nennen verschiedene Autoren von 19 bis 23 typische Konsonanten/4/. L.N.Charitonow /5/ zählte in der jakutischen Sprache 20 Konsonantenphoneme und teilte sie in Geräuschphoneme und sonore Phoneme ein; Geräuschphoneme werden weiter in stimmlose und stimmhafte eingeteilt. P.P. Baraschkow nennt im Jakutischen 27 Konsonantenphoneme /6/, E.I. Ubrjatowa – 21 /7/, N.D. Djatschkowskij – 20 /8/.

Die Bestimmung des Bestandes der Konsonantenphoneme hat der Autor auf drei Etappen verwirklicht (im Laufe der Arbeit haben sich diese Etappen gemischt).

Auf der ersten Etappe wurden alle Konsonanten schriftlich festgelegt, welche während des Aufschreibens der Texte, einzelner Wörter und beim Lesen der Texte, die die anderen Expeditionsteilnehmer aufgeschrieben haben, fixiert wurden, was man eigentlich auch als Gehöraufnahme bezeichnen kann. Man stellte im Dolganischen folgende Konsonanten fest, die durch das Konsonantensystem der russischen Sprache aufgenommen wurden: б , б , п , п , м , м , д , д̧ , т , т , н , н , с , с , р , р , л , ł , л’, нь, дь, дт , чь , ть, j , ң , г , к , ҕ , k , kʰ, ħ , ɣ , h , ɦ (nach der Transkription von A.P. Dulson).

Die zweite Etappe – die Prüfung des Konsonantenbestandes mittels Analyse der aufgeschriebenen Texte (Vergleich der Wortformen eines Lexems und der der verwandten Lexeme mit verschiedener Semantik, das infolgedessen einen vollen Konsonantenbestand im Dolganischen gegeben hat. Unten ist der Prozeß der Arbeit auf der zweiten Etappe dargestellt (in der Tat war es viel komplizierter und nicht so systematisch).

Beim Vergleich der Wortformen mit Generabedeutung masta: "hack Holz" /9/ - mastan "versorge dich mit Holz", mastat "bitte, daß man dir Holz hackt", mastas "hack Holz mit jemand zusammen" stellt man 3 Konsonanten - n, t, s fest, die in diesen Wörtern mit den Generabedeutungen der Reflexivität, Kausalität und Kooperativität zu beziehen sind.

Die Gegenüberstellung der Gleichstammlexeme masta : "hack Holz" - masta:r "hack Holz später", masta:q "derjenige, der Holz hat" macht uns die Konsonanten "r" und "q" bekannt, der erste von ihnen ist die obligatorische Komponente des Morphems - a:r mit der grammatischen Bedeutung des zukünftigen Imperativs, die zweite - die obligatorische Komponente des Morphems +ta:q des wortbildenden Postfixes der Adjektive des Bezitzens. In den Wortformen maspɐn "mein Holz" (Akk), maskɐn "dein Holz" (Akk) sieht man die Konsonanten p und k, der erste - im Morphem des Akkusativs der Personalpossesiven Deklination der ersten Person Sg., der zweite - im Morphem des Akkusativs der personalpossesiven Deklination der zweiten Person Sg. Das Vorhandensein des Lautes k wird auch mit dem Stamm des Adjektivs ilimnæ:k "derjenige, der ein Netz hat" bei der Gegenüberstellung mit der verbalen Wortform ilimnæ: "fang Fische mit einem Netz" bestimmt.

Die Wortformen tabam "mein Hirsch" und tabaŋ "dein Hirsch" mit grammatischen Bedeutungen der persönlichen Angehörigkeit heben bei der Gegenüberstellung miteinander und auf Grund des Ausgangsstamms taba "Hirsch" die Komponenten m und n hervor, dementsprechend sondern die objektiven Personalpossesivformen dieses Lexems tababɐn "meinen Hirsch" (Akk), tabagɐn "deinen Hirsch" (Akk), die Konsonanten b und g aus.

Aus dem Vergleich der Lexeme ʧaj "Tee" - ʧajda "Tee", ʧaja "sein Tee" sondert man den Konsonanten d aus, als eine obligatorische Komponente des Postfixes da im Vergleich zu dem Postfix der Personenangehörigkeit der dritten Person +a.

Aber nicht alle Konsonanten nehmen an den für ihre Vergleichabsonderung bequemen Positionen teil; in solchen Fällen benutzt man die Methode der Gegenüberstellung von nahlautenden Wurzelstämmen mit verschiedener lexischen Bedeutung (die Methode der Pseudohomonyme). Auf solche Weise hebt man Konsonanten hervor, ohne die die Lauthülle des gegebenen Lexems mit ihrer lexischen Bedeutung nicht unterstützt wird. Der Vergleich von Wörtern aʧa "Darm", ala "Träger", aɣa "Vater", ajɨ: "heiliger Anfang", ajɨ "Sünde", ɲon "Volk", hon "Oberbekleidung", ɛbɛ "Großmutter", ɛʰɛ "Großvater" läßt uns die Laute ʧ, l, ɣ, j, ʃ, ɲ, h, ɦ anführen. Genauso werden das harte l und das weiche ʎ beim Vergleich der eingstammigen Lexeme bult "Jagd" und bulʎɐt "Jäger" hervorgehoben.

Die auf der zweiten Etappe der Forschung festgelegten 25 typischen Konsonanten wurden für die Bequemlichkeit ihrer künftigen Analyse in 6 Gruppen nach dem aktiven Organ der Artikulation vereint: Lippen - p, b, m ; Vorderzungen - t, d, n, s, l, r ; Mittelzungen - ʧ, ʤ, ɦ, ɲ, j, ʃ, ʎ ; Hinterzungen - k, g, ɣ, ŋ ; Zäpfchen - q, ʒ, ɣ ; Rachenlaute - h, ɦ .

Auf der dritten Etappe wurde der Bestand der Konsonantenphoneme der dolganischen Sprache bestimmt, hauptsächlich nach den Regeln der Aussonderung der Phoneme von N. Trubezkoj /10/ mit Analyse der morphologischen Struktur der Wortformen, wo es notwendig war. Auf solche Weise werden in jeder Artikulationsgruppe folgende Konsonantenphoneme ausgesondert: Lippen - [p]1, [p]2, [m] ; Vorderzungen - [t]1, [t]2, [s], [n], [l], [r]; Mittelzungen - [ʧ]1, [ʧ]2, [ɲ], [j] ; Hinterzungen - [k]1, [k]2, [ŋ] ; Zäpfchen - [q]1, [q]2 ; Rachenphoneme - [h]. Die Ergebnisse der weiteren Analyse zeigen, daß die Hinterzungen - [k]1, [k]2, die Zäpfchenphoneme [q]1 und [q]2 im Verhältnis der zusätzlichen Distribution zueinander stehen. Das läßt uns eine weitere Phonemverallgemeinerung machen, indem die beiden Phonempaare [k]1, [k]2 und [q]1, [q]2 zu einem Paar vereint werden, auf die bedingt die Symbole [k]1 und [k]2 aber mit bedeutend verallgemeinerterem Inhalt verbreitet werden. Im allgemeinen unterscheidet man im Dolganischen 17 Konsonantenphoneme.

Der Autor hält sich an jene Auffassung des Phonems, die von L.W.Tscherba /11/ formuliert und von L.R.Sinder /12/ in unserer Zeit entwickelt wurde.

Die durchgeführte experimental-phonetische Forschung läßt folgende Schlußfolgerungen ziehen.

Für das System der Konsonantenphoneme der dolganischen Sprache ist die Einteilung in zwei phonetische Gruppen nach dem Spannungsgrad des aktiven Hauptorgans eigen: schwache (Geräuschlaute) - 10 Phoneme und sehr schwache (geräuscharme Laute) - 7 Phoneme.

Die schwachen Phoneme werden ihrerseits in lange (5 Phoneme) und kurze (5 Phoneme) eingeteilt.

Die Tönungen der ersten Phoneme in den folgenden Paaren [p]1 - [p]2, [t]1-[t]2, [ʧ]1-[ʧ]2, [k]1-[k]2 werden vom Forscher, dessen Muttersprache Russisch ist, als stimmhafte, und die Tönungen der nächsten Phoneme als stimmlose Laute aufgenommen, obwohl in einzelnen Tönungen der ersten Phoneme die stammhafte Komponente nach experimentalen Angaben zwischen 34,2 - -100,0% der gesamten Lautlänge, und in den Tönungen der nächsten Phoneme die stimmhafte Komponente zwischen 0,0-35,7% der gesamten Lautlänge schwankt. Aber das Vorhandensein solcher stimmhaften wie «ɦ» und stimmlosen wie «h» der kombinatorischen Positionstönungen beim Phonem [h] stört die eben bemerkte Gesetzmäßigkeit in der phonetischen Gegenüberstellung der dolganischen Konsonanten nach den Merkmalen der Stimmlosigkeit - Stimmhaftigkeit. Was aber die Gegenüberstellung nach der Länge anbetrifft, so ist sie in der Tat, wie die experimentalen Angaben zeigen, ausnahmlos, indem sie alle Geräuschkonsonanten in 2 Gruppen einteilt - in kurze und lange: [p], [t], [ʧ], [k], [h] ; [p:], [t:], [ʧ:], [s:], [k:].

Zum Beispiel haben die Tönungen des Phonems [t] in der intervokalen Position =VCV= eine relative Länge (nach verallgemeinerten Angaben von 3 Sprechern) etwa 38,0-108,0% der mittleren Länge des Lautes, und die Tönungen des Phonems [t:] in derselben Position =VCV= eine relative Länge von etwa 96,3-168,0% der mittleren Länge des Lautes.

Die Teilübereinstimmung der Zonen in konkreten Schwankungen der relativen Längen von Tönungen kurzer und langer Geräuschphoneme läßt sich bei der exakten Zonenverteilung ihrer relativen Mittellängen dadurch erklären, daß Dolganisch, welches der Jakutischen Sprache nahverwandt ist, sich als Sprache in einer verhältnismäßig kurzen historischen Frist, ungefähr im Laufe von 3 Jahrhunderten, und dabei unter komplizierten etnogenetischen Bedingungen bei Teilnahme verschiedener Gruppen der türkischen, tungußermantschurischer und samojedischer Sprachfamilien herausbildete.

Für alle Geräuschkonsonantenphoneme der dolganischen Sprache ist die Mundartikulation charakteristisch; die Mundartikulation ist das obligatorische Hauptmerkmal dieser Gruppe von Konsonantenphoneme.

Die Analyse der übereinstimmenden Dentopalatogramme der Tönungen der Phoneme [t]// [t:], [ʧ] // [ʧ:] zeigt nur einen geringen Unterschied in der Form des Abdrucks auf dem künstlichen Gaumen, daß die Muskelspannung bei der Artikulation der inneren stimmlosen Verschlußkomponente in den Tönungen der Phoneme [t:] [ʧ:] sich nicht wesentlich von der Muskelspannung der inneren Verschlußkomponente in den Tönungen der Phoneme [t], [ʧ] unterscheidet. Die Vorderzungenkonsonantenphoneme [t] und [t:] und die Mittelzungenphoneme [ʧ] und [ʧ:] werden also nach Stärke und Schwäche nicht gegenübergestellt.

Der Vergleich von Ergebnissen der intervokalen homorganen Phoneme [p], [p:] ; [t], [t:] ; [ʧ], [ʧ:] ; [k], [k:], die wir mittels eines Oszillographen bekommen haben, zeigt, daß auf der Mundlinie die Segmente dieser Konsonanten mehr oder weniger Engekomponenten haben - minimal für die Tönungen der Phoneme [p:], [t:], [ʧ:], [k:], etwas mehr in den Segmenten der Phoneme [p], [t], [ʧ], [k], die neben den Engeabschnitten auch schwache Verschlußabschnitte haben, die intervokalen Tönungen der Phoneme [p], [t], [ʧ], [k] können nur schmale Engen haben, aber der Experimentator unterscheidet nicht vom Gehör die Konsonanten mit dieser schmalen Engeartikulation von Konsonanten mit Verschlußartikulation, wobei er sie miteinander vermischt, darum werden in der vorliegenden Arbeit diese Konsonanten als Verschlußlaute bezeichnet.

Diese fakultative Enge in den Geräuschphonemen zeugt undirekt von der verhältnismäßig schwachen Muskelspannung der aktiven Organe bei ihrer Artikulation. Eine verhältnismäßig größere Engeartikulation in den Segmenten der kurzen Phoneme [p], [t], [ʧ], [k] im Vergleich zu den Segmenten der langen Phoneme [p:], [t:], [ʧ:], [k:] läßt sich nicht nur durch größere Spannung der letzten erklären, obwohl auch das möglich ist, sondern auch durch temporale Bedingungen, welche die Verwirklichung dieser ihrem Wesen nach Verschlußartikulation der langen Verschlußphoneme sichern. Die analysierten Angaben der experimentalen Forschung lassen behaupten, daß die mögliche Gegenüberstellung der homorganen Konsonanten nach dem Spannungsgrad des aktiven Organs, die den anderen türkischen Sprachen eigen ist, für das dolganische Konsonantensystem irrelevant ist.

Die Aspiration oder ihr Fehlen können auch nicht als Hauptmerkmale dienen, weil sich einerseits vom Gehör die Verschlußlaute der dolganischen Sprache von den russischen Verschlußkonsonanten nicht unterscheiden und andererseits im experimentalen Stoff, welchen man mit Hilfe eines Oszillographen bekam, die für die aspirierten Konsonantenkomponenten charakteristischen Abschnitte nur in zwei Positionen entdeckt wurden: im Auslaut und, in einigen Fällen, auch im absoluten Anlaut, und dabei kann ein und derselbe Sprecher den Laut entweder schwach aspiriert oder gar nicht aspiriert aussprechen.

Zehn dolganische Geräuschkonsonanten haben eine verschiedene Positionsverteilung in der Lautstruktur des Wortes. Außer einem kommen alle langen Geräuschphoneme in einer beliebigen Position vor: im An-, In- und Auslaut; das lange Phonem [ʧ:] wird im Auslaut nicht gebraucht. Außer dem Auslaut werden alle kurzen Geräuschphoneme in den Positionen =CV= und =VCV= gebraucht.

Die Phoneme [m], [n], [l], [r], [ɲ], [j], [ŋ] kommen hauptsächlich in stimmhaften Tönungen vor, manchmal aber auch in stimmlosen und zum Teil stimmlosen Tönungen. Infolgedessen werden die genannten

Phoneme nicht als sonore, sondern als sehr schwache geräuscharme Phoneme bestimmt. Gewöhnlich sind die obengenannten Merkmale miteinander verbunden; dabei wird das zweite Merkmal vom ersten bedingt. Für die dolganischen Laute m, n, ɲ, ŋ, l, r, j ist wirklich, wie es vom Gehör bestimmt wird, ein kleinerer Geräusch im Vergleich zu den Phonemen, die als Geräuschphoneme bezeichnet werden, charakteristisch, deshalb kann man diese Gruppe der Phoneme als sehr schwache geräuscharme Phoneme nennen.

Der Autor meint, daB es zweckmäBig wäre, die geräuscharmen Phoneme der dolganischen Sprache in zwei Gruppen einzuteilen: Nasale-und Mundphoneme.

Für die nasalen geräuscharmen Phoneme [m], [n], [ɲ], [ŋ] ist in artikulatorischer Hinsicht folgendes typisch: a) ein sehr schwach gespannter VerschluB in der Mundhöhle, der in gewissem MaBe durch einen EngeverschluB ergänzt, sehr oft auch durch den letzten ersetzt wird, was nur bei schwacher Muskelspannung des aktiven Organs möglich ist; b) der Abgang des weichen Gaumens von der hinteren Wand des Pharyn's.

Die geräuscharmen Mundkonsonanten l, r, j vereint artikulatorisch folgendes: a) Enge in der Mundhöhle b) VerschluB des weichen Gaumens mit dem hinteren Teil des Pharynx's, wobei die Luft durch die Nasenhöhle nicht strömen kann.

Diese drei Phoneme unterscheiden sich voneinander durch die Arten von Hindernissen mit entsprechenden akustischen Effekten - Artikulation der Seitenenge (Engen) beim Phonem [l], der Mittelzungen-und Vordergaumenenge bei [j], Mittelvorderzungealveolarenge bei [r].

Die Konsonantenphoneme der dolganischen Sprache werden nach dem Hauptorgan und den entsprechenden passiven Organen in fünf phonematische Gruppen eingeteilt: 3 Phoneme der ersten Artikulation (Lippenphoneme), 6 Phoneme der zweiten Artikulation (Vorderzungenphoneme), 4 Phoneme der dritten Artikulation (Mittelzungenphoneme) 3 Phoneme der vierten Artikulation (Hinterzungen-und Ovularphoneme), 1 Phonem der fünften Artikulation (Rachenphonem).

Die Hauptmerkmale der artikulatorisch-akustischen Basis der dolganischen Sprache auf dem Gebiet der Konsonanten sind folgende:

1) Bei der Aussprache der Konsonanten werden die Sprachorgane verhältnismäBig wenig oder kaum gespannt.

2) Für den Sprechapparat ist das Vorhandensein von einfachen und kombinierten Artikulationen eigen. Ein besonders breiter Diapason ist nach dem aktiven Organ den Konsonanten der vierten Artikulation eigen. 3) Das Vorhandensein von 7 Artikulationsreihen der Konsonanten: a) Lippenkonsonanten, b) Vorderzungenalveolar-oder Dentalalveolar-hauptsächlich Dorsallaute, c) Mittelzungen-, Alveolar-und Vordergaumenlaute, d) Hinterhartengaumenlaute, e) Vorderweichgaumenlaute, f) Hinterzungen-Hintergaumenlaute, g) Rachenlaute.

4) Es gibt wenig Zisch-, besonders Geräuschlaute.

Literatur

/I/ Е.И.Убрятова, О языке долган, "Языки и фольклор народов Сибирского Севера", М.-Л., 1966.

/2/ "Народы Сибири", М.-Л., 1956. Nikolaus Poppe, Das Jakutische, "Philologiae Turciacae Fundamenta", t.1. Wiesbaden, 1959; Karl Menges, "The turcic languages and peoples", Wiesbaden, 1968.

/3/ Е.И.Убрятова, "Язык норильских долган", Изд-во "Наука" СО АН СССР, 1985.

/4/ Otto Böhtlingk. "Über die Sprache der Jakuten". St.Petersburg, 1851; С.В. Ястремский. "Грамматика якутского языка" Иркутск, 1900; W. Radloff, "Die jakutische Sprache in ihrem Verhältnisse zu den Türksprachen". St.Petersburg, 1908; "saqalɯ: bicik", Якутск, 1917; Н.Н.Поппе. "Учебная грамматика якутского языка". М., 1926.

/5/ Л.Н.Харитонов, "Современный якутский язык", Якутск, 1947.

/6/ П.П.Барашков. "Звуковой состав якутского языка", Якутск, 1953.

/7/ Е.И.Убрятова. Якутский язык. "Языки народов СССР".

/8/ Н.Д.Дьячковский. "Звуковой строй якутского языка". Якутск, 1977.

/9/ В.М.Наделяев, "Проект универсальной фонетической транскрипции", М.-Л., 1960.

/10/ Н.С.Трубецкой, "Основы фонологии", М., 1960.

/11/ Л.В.Щерба. "Языковая система и речевая деятельность", Л., 1974.

/12/ Л.Р.Зиндер, "Общая фонетика", Л., 1960.

ФОНОЛОГИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ЗВУКОВ "СЛАБОЙ" ПОЗИЦИИ /НА МАТЕРИАЛЕ БЕЗУДАРНЫХ ГЛАСНЫХ ФРАНЦУЗСКОГО ЯЗЫКА/

АКЧЕКЕЕВА МАРИЯ СУЛТАНОВНА

Кафедра французского языка
Киргизский государственный университет
Фрунзе, Киргизия, СССР 720024

Работа посвящена проблеме фонемной интерпретации безударных гласных среднего подъема французского языка, а также анализу соотношения фонологического восприятия и акустических характеристик звуков. Опираясь при определении фонологического статуса безударных гласных на речевое поведение носителей языка, предлагается их однозначная фонемная трактовка; выясняются факторы, действующие на использование фонем в безударных слогах, а также роль в этом более высоких уровней языкового строя.

Одной из наиболее сложных фонологических проблем для многих языков, в том числе и французского, является фонемная трактовка звуков, находящихся в так называемой "слабой" позиции, т.е. в позиции, где отсутствуют некоторые противопоставления фонем. Применительно к французскому языку это относится, в основном, к определению фонемной принадлежности гласных /e-ɛ/, /ø-œ/, /o-ɔ/ в безударных слогах. Фонологическая трактовка этих гласных у разных исследователей неодинакова, что связано с различным подходом к решению ряда общефонологических проблем /нейтрализация фонемных противопоставлений, чередование фонем, типы произнесения и стили произношения/. Существует мнение, согласно которому противопоставление гласных среднего подъема в безударной позиции нейтрализуется, причем использование той или иной фонемы подчиняется гармонии гласных по подъему, т.е. обусловлено качеством гласного последующего ударного слога /I /. Данная точка зрения восходит к М.Граммону /2/.

Другие полагают, что противопоставления фонем по признаку подъема в этой позиции сохраняются, так как невозможно объяснить все случаи дублетного произношения слов гармонией гласных /3/.

Согласно третьей точке зрения, сущесвуют "средние" гласные, промежуточные по своему качеству между открытыми и закрытыми. Такие формулировки принадлежат авторам-дофонологам или же авторам, не решающим фонологических проблем /4/. Л.В.Щерба допускает средние гласные лишь в небрежном произношении; в отчетливом произнесении они проясняются в открытые или закрытые /3/. Следовательно, по Л.В.Щербе, фонемная трактовка звуков должна опираться на отчетливое произношение.

М.Граммон, не принявший фонологических идей, резко возражал против средних гласных, очевидно, потому, что его непосредственное восприятие носителя языка требовало отнесения звука к той или иной четко определенной звуковой единице -фонеме /2/.

Авторы, исследовавшие объективные характеристики безударных гласных среднего подъема во французском языке, не обращались к их восприятию носителями языка, что позволило бы решить вопрос об их фонемной интерпретации. Принадлежность гласного к той или иной фонеме считалась ими заранее заданной.

Таким образом, для решения спорных фонологических вопросов необходимо обращение к широкому аудиторскому анализу наряду с исследованием акустических характеристик соответствующих звуков.

Расхождения между разными словарями, между словарями и рекомендациями орфоэпических пособий, несовпадение тех и других данных с фонологическими описаниями, в которых содержатся утверждения о нейтрализации некоторых оппозиций в безударном слоге, требуют специального анализа употребления гласных.

В качестве материала для такого анализа были использованы данные Словаря А.Мартине и А.Вальтер, не предписывающего, а описывающего произносительную норму /5/.

На основе статистической обработки этих данных были выявлены закономерности использования носителями языка тех или иных форм произношения. Подсчитывались все случаи разного выбора информантами гласных /e-ɛ/, /ø-oe/, /o-ɔ/ в безударных слогах. Учитывались, кроме того, различные буквенные обозначения этих гласных, а также другие факторы, способные влиять на произношение безударных гласных: качество гласного ударного слога, наличие или отсутствие соответствующего ударного корня, тип слога.

Употребление гласных фонем среднего подъема в безударных слогах не может быть описано простыми жесткими правилами ни для группы говорящих, ни даже для одного носителя языка. Поэтому нельзя говорить и о нейтрализации противопоставлений по подъему. В современном литературном языке допустима достаточно широкая вариативность фонемного состава слова.

Анализ произношения каждого из информантов позволил говорить о преимущественном использовании открытой или закрытой фонемы: в паре негубных гласных переднего ряда предпочтительным для большинства дикторов является в открытом безударном слоге фонема /e/, перед сочетанием согласных - /ɛ/; в паре гласных заднего ряда в безударной позиции преимущественно используется фонема /ɔ/; для пары /ø-oe/ в открытом слоге характерно несколько более частое употребление фонемы /oe/ перед слогом с открытым гласным, /ø/ - перед слогом с закрытым гласным; перед сочетанием согласных преимущественно используется форма с /ø/. Таким образом, анализ данных Словаря позволяет говорить о некоторой тенденции к гармонии по подъему лишь для последней пары гласных.

Для выбора из двух фонем в безударной позиции имеют значение и другие обстоятельства: важную роль играет морфологическая аналогия, т.е. распространенным является произношение того гласного в безударном слоге слова, который звучит в соответствующем ударном корне. Существенным является влияние орфографии на произношение.

В то же время то обстоятельство, что фонетические словари не фиксируют "средних" гласных, произношение которых описано в ряде исследований и пособий, заставляет поставить вопрос: в какой мере отраженная в словарях картина соответствует реально произносимым звукам? Как фонемная классификация, проведенная на основании восприятия слова в целом, соотносится с акустическими характеристиками звуков и с их восприятием вне смыслового контекста? Для выяснения этих вопросов был предпринят аудиторский эксперимент и исследованы объективные спектральные характеристики гласных . Восприятие гласных анализировалось в два этапа: I/ восприятие гласных, выделенных из изолированных слов и словосочетаний; 2/ восприятие гласных, выделенных из фраз, в состав которых вошли те же самые изолированные слова. Всего было подобрано 555 изолированных слов и словосочетаний, из них 92 были введены в состав фраз. Гласные были представлены во всех возможных позициях, при этом учитывались различные факторы, способные влиять на выбор фонемы безударного слога.

Для того, чтобы дикторы не сравнивали при чтении буквенные обозначения, что могло им подсказать то или иное произношение, списки слов, включавшие различные гласные, были составлены в случайном порядке. Гласные, выделенные из этих слов, также в случайном порядке предъявлялись на опознание аудиторам. Каждый гласный, выделенный из контекста при одновременном слуховом и визуальном контроле, был переписан на чистую ленту и повторялся три раза. Затем следовала небольшая пауза длительностью в несколько секунд, во время которой аудиторы записывали свой ответ. В качестве задания было предложено обозначить тот гласный французского языка, который они слышат. При этом заранее предполагалось, что ими могут быть восприняты гласные, которых нет в эксперименте.

Гласные, выделенные из изолированных слов, словоформ и словосочетаний, в произношении каждого диктора составили I6 серий для прослушивания по 35 стимулов в каждой; две серии по 3I стимулу и одна серия с 30 стимулами содержали гласные, выделенные из фраз.

В опытах по восприятию участвовало 20 человек, жителей различных городов Франции, из них половина прослушала гласные в произношении диктора I, остальные - гласные, произнесенные диктором 2. Всего по двум дикторам обработано I2920 ответов.

Известно, что при выделении гласного мы лишаем его тем самым всяких указаний на то, что определенные акустические свойства зависят от фонетического контекста. Это в дальнейшем повлияло на восприятие:так, гласные /e-ɛ/ воспринимались как соответствующие губные в положении между губными согласными, а гласные /o-ɔ/ - как соответствующие гласные переднего ряда в положении между переднеязычными согласными. Но поскольку главной задачей исследования был анализ восприятия признака подъема, названные особенности восприятия не мешали интерпретации результатов аудиторского анализа.

Анализ результатов восприятия был проведен с помощью критерия знаков. Если из десяти аудиторов семь или более указывали на закрытость гласного, это считалось достаточным основанием для утверждения, что полученный результат не случаен, т.е. что гласный в данном конкретном слове оценивался как закрытый. Соответственно гласный считался открытым, если не менее семи человек опознавали его как открытый. Но если около половины аудиторов /четыре, пять или шесть человек/ указывали на открытость гласного, а остальные свидетельствовали о его закрытости, то восприятие признака подъема такого гласного считалось случайным.

Случайное восприятие, связанное с неуверенностью аудиторов в своем выборе, указывает на средние физические характеристики гласного. Иными словами это то, что иногда фонетисты называют "средний гласный". Для обозначения гласных, характеризующихся случайным опознанием признака подъема, в дальнейшем используется термин "средний", а также применяются знаки: E - для среднего гласного из пары /e-ɛ/, OE - для гласного из пары /ø-oe/, O -для среднего гласного из пары /o-ɔ/.

На основе анализа восприятия гласных /e-ɛ/ был получен вывод о преимущественном использовании /e/ в открытом безударном слоге, /ɛ/ - перед сочетанием согласных.

Произношение /ɛ/ в открытом безударном слоге можно объяснить влиянием морфологической аналогии: у первого диктора 77% слов, где безударный гласный был воспринят аудиторами как открытый, имеют /ɛ/ в соответствующих ударных корнях: у диктора 2 такие случаи составляют 90%. В ряде слов произношение открытого гласного невозможно объяснить морфологической аналогией, но тогда очевидным является влияние орфографии на произношение: /ɛ/ произносится на месте орфографических ai, aî, ê, ei.

Однако нередки случаи, когда, несмотря на возможное влияние морфологической аналогии и орфографии, гласный все же произносится как закрытый: guêpière, (ils) vinaigrèrent /д.I/, faiblesse, aigrette /д.2/ с /e/ в безударном слоге. Подобные случаи свидетельствуют о наличии чередования /ɛ/ ударного корня с /e/ безударным. Но возможность произнесения открытого гласного в безударном слоге по аналогии с ударным /ɛ/ или под влиянием орфографии делает это чередование необязательным, факультативным.

Аудиторский анализ гласных /o-ɔ/ показал, что в произношении д.I чаще опознается /ɔ/ в безударной позиции, у д.2 - /o/. Закрытый гласный произносится, как правило, на месте орфографических au, eau, ô, а открытый - на месте буквы o, что связано с историческими причинами. Как и в предшествующей паре, существенную роль при выборе фонемы безударного слога играют морфологические отношения. Гармония гласных здесь также не находит подтверждения, что относится и к третьей паре гласных /ø-oe/. Эксперимент по восприятию гласных /ø-oe/ выявил существование определенной зависимости выбора безударной фонемы от позиции: /oe/ опознается преимущественно перед сочетанием согласных, т.е. в позиции, в которой возможен закрытый слог, /ø/ - в открытом слоге.

Гласные со случайно воспринятым признаком подъема зафиксированы во всех трех парах, хотя и в очень незначительном числе случаев. Процент средних гласных несколько выше во фразах, чем в изолированных словах, что может свидетельствовать о меньшей четкости звуков, произнесенных во фразах. У средних гласных случайно опознавался не только подъем, но и другие признаки /ряда и огубленности/.

Традиционные, широко распространенные взгляды, связывающие открытость или закрытость гласного с качеством ударной фонемы не подтверждаются в результатах исследования: зависимости выбора фонемы безударного слога от качества гласного последующего ударного слога не существует. Фонемное противопоставление гласных сохраняется в безударной позиции, при этом существует преимущественное использование фонем в той или иной позиции: I) для гласных переднего ряда характерно преимущественное употребление закрытых /e, ø/ в открытом слоге, открытых /ɛ, oe/ - перед сочетанием согласных; 2) для гласных заднего ряда можно говорить о зависимости варьирования фонем от индивидуальных особенностей дикторов /для д.I предпочтительна форма с /ɔ/, для д.2 - форма с /o/.

Таким образом, несмотря на внешнюю симметричность этих трех пар гласных /их объединяет противопоставленность по признаку подъема/, с точки зрения функциональной они не аналогичны, что связано с различной дистрибуцией гласных переднего и заднего ряда. В паре гласных заднего ряда ограничение, связанное с позицией в слове, имеет фонема /ɔ/, в паре негубных гласных дистрибутивно ограничена открытым слогом фонема /e/. Именно эти фонемы оказываются предпочтительными в безударной позиции.

Для выяснения вопроса, чему объективно соответствует средний гласный и что затрудняет восприятие признака открытости-закрытости, был проведен спектральный анализ гласных. Было снято I290 спектрограмм типа "Видимая речь". Исследовались спектральные характеристики стационарной части гласного, измерялись частотные значения F_I и F_2, а также учитывалось движение F_2 в тех случаях, когда формантные переходы составляли его значительную часть /не менее 40% длительности/.

В результате проведенного анализа было обнаружено, что восприятие той или иной фонемы достаточно определенным образом связано с объективными акустическими характеристиками звука. При отсутствии какой-либо информации смысла аудитор принимает решение о фонемной принадлежности звука на основании только звучания самого гласного. В табл. I приведены средние значения формант безударных гласных, в том числе и гласных со случайно воспринятым признаком подъема.

Таблица I.

Средние значения F_I и F_2 безударных гласных /в Гц/

гласные фонемы	диктор I		диктор 2	
	F_I	F_2	F_I	F_2
/e/	445	2000	450	2290
/ɛ/	530	I850	525	I970
E	475	I860	470	I900

/ø/	450	1550	455	1490
/oe/	520	1390	525	1580
OE	470	1500	475	1420
/o/	450	1010	450	980
/ɔ/	540	1250	535	1100
O	485	1130	475	1070

Случайное восприятие признака подъема связано, во-первых, со средним качеством самого гласного /F_I такого гласного, названного нами "истинно средним", равна 485-500 Гц/, а во-вторых, с влиянием соседних согласных, чему объективно соответствует движение F_2 гласного.

Однако в ряде случаев такое восприятие оказалось необъяснимым с точки зрения спектральных характеристик из-за совпадения с формантными значениями закрытых, реже открытых, гласных. Так, безударный гласный в словах désarmer, (ils)rêvèrent, férié /д.2/ был воспринят аудиторами как закрытый при F_I=450 Гц, F_2=1950 Гц, а в словах testacé, c'est ici, c'est assez, c'est-à-dire при тех же значениях формант был воспринят случайно.

Анализ количественных характеристик средних гласных показал, что их длительность в 1,5-2 раза меньше длительности гласных, воспринятых как открытые или закрытые /табл.2/.

Таблица 2.

Зависимость восприятия признака подъема от длительности гласных

вос-при-нято	F_I	F_2	средняя длительность в мсек
/e/	440-460	1950	100
E	450-460	1950	65
/ɛ/	485	1950	100
E	485	1950	80

Таким образом, одни лишь формантные характеристики не обеспечивают уверенной фонемной идентификации, для которой необходима еще и достаточная длительность гласного. Акустические характеристики средних гласных объясняют, почему эти звуки чаще других характеризуются случайным восприятием не только подъема, но и других признаков. Восприятие того и другого признака зависит от положения F_2. У средних E F_2 либо целиком сдвигается вниз, либо имеет длительный переходный участок с более низким положением этой форманты. У средних O, напротив, F_2 целиком сдвинута кверху или же имеет более высоко расположенный переход. В отсутствие информации о том, что такая особенность звучания вызвана фонетическим положением, и при дефиците времени аудитор принимает случайное решение о соответствующем признаке гласного.

В результате проведенного исследования было обнаружено, что фонемное противопоставление по признаку открытости-закрытости гласных среднего подъема во французском языке сохраняется в безударной позиции. Предпочтительными являются те фонемы, которые имеют дистрибутивные ограничения.

Хотя в системе языка гласные пар /e-ɛ/, /ø-oe/, /o-ɔ/ противопоставлены по одному и тому же признаку, дистрибутивные ограничения у них различны, и поэтому в одной и той же фонетической позиции предпочтительными оказываются гласные с различными дифференциальными признаками /закрытый в паре негубных и открытый в паре гласных заднего ряда/. Это обстоятельство следует иметь в виду при рассмотрении системных отношений между фонемами и анализе их функционирования. Иными словами, определение дифференциальных признаков недостаточно для того, чтобы описывать функциональные отношения в системе.

Анализ употребления гласных среднего подъема в безударных слогах выявил ряд закономерностей, которым подчиняется выбор орфоэпических вариантов у говорящих. Эти закономерности определяются не только фонетически, но и /в тех случаях, где это возможно/ смысловыми и морфологическими отношениями. Наличие той или иной фонемы в ударном корне способствует сохранению ее и при безударной позиции того же слога. Следовательно, при анализе собственно фонетических закономерностей следует принимать во внимание и более высокие уровни языкового строя, т.е. морфологические отношения.

Качество гласных опознается достаточно уверенно, причем именно подъем, т.е. тот признак, по которому, как часто утверждают, нейтрализуются эти оппозиции, опознается лучше, чем другие признаки. Термин "нейтрализация" не может быть интерпретирован как ограниченность употребления фонем, потому что, как показывают и анализ фонетических источников, и исследование восприятия, в любой позиции возможна любая из двух фонем.

Неуверенное восприятие признака подъема наблюдалось только в тех случаях, когда гласный обладал малой длительностью, и лишь в очень небольшом числе случаев такие средние гласные имели и среднее значения F_I, т.е. могли рассматриваться как качественно редуцированные /что зависело от их удаленности по отношению к ударению и также было связано с их краткостью/. Таким образом, средние гласные практически отсутствуют в исследованном материале. Восприятие же средних гласных связано с количественной, либо – реже – с качественной редукцией, т.е. с неполным типом произнесения.

/1/ А.Мартине,"Нейтрализация и синкретизм", Вопросы языкознания, 1969.
/2/ M.Grammont, "Traité pratique de prononciation française", Paris, 1951.
/3/ Л.В.Щерба,"Фонетика французского языка", Москва, 1955.
/4/ A.Lombard,"Remarque sur le e moyen du francais. Mélanges de linguistique offerts a Dauzat", Paris, 1951.
/5/ A.Martinet, H.Walter,"Dictionnaire de la prononciation française dans son usage réel", Paris, 1973.

ВТОРИЧНЫЕ ТИПЫ СЛОГОВЫХ ИНТОНАЦИЙ В ЛИТОВСКИХ ДИАЛЕКТАХ

АЛЕКСАС ГИРДЕНИС

Вильнюсский госуниверситет
Кафедра литовского языка
Вильнюс, ЛитССР, СССР 232055

ГЕНОВАЙТЕ КАЧЮШКЕНЕ

Шяуляйский пединститут
Кафедра литовского языка
Шяуляй, ЛитССР, СССР 235419

0. Резюме. В докладе приводятся экспериментальные данные, свидетельствующие о существовании в литовских диалектах двух вторичных типов слоговых интонаций, выполняющих различительную функцию как в акутированных, так и в циркумфлектированных слогах.

I.I. В северожемайтских диалектах литовского языка недавно отмечено существование двух вторичных типов слоговых интонаций (акута и циркумфлекса), противопоставляемых в одинаковых фонетических условиях /I/. Можно выделить следующие основные случаи их функционирования:

а) "баритонические" (I и 2 акцентная парадигма) и "окситонические" (т. е. подвижные; 3 и 4 а. п.) именные части речи, ср.: kâ·ìts$_{(1)}$ '(сущ.) долото' : ka.ȋts$_{(2)}$ '(прич.) бит(ый)', svệ·ists$_{(1)}$ '(сущ.)(сливочное) масло' : svẹ.ȋsts$_{(2)}$ '(прич.) кинут(ый)', sã·usi$_{(1)}$ '(вин. п.) январь' : sã.u.si$_{(2)}$ '(вин. п.) тлю';

б) глаголы одноразового и многократного действия, ср.: trâ·ukẹ$_{(1)}$ 'тянул(а)' (инф. trâ·uktẹ) : tra.ûkẹ$_{(2)}$ 'дергал(а)' (инф. tra.úkî·tẹ), brã·ukẹ$_{(1)}$ 'перечеркнул(а)' (инф. brã·uktẹ) : brã.u.kẹ$_{(2)}$ 'перечеркивал(а)' (инф. braukî·tẹ);

в) 3 лицо настоящего и будущего времени, напр.: наст.в. kâ·iš$_{(1)}$ 'чистит (-ят)' : буд. в. ka.îš$_{(2)}$ 'будет (-ут) чистить', наст. в. mẹ̃·iš$_{(1)}$ 'смешивает (-ют)' : mẹ̃.i.š$_{(2)}$ 'будет (-ут) смешивать'.

Приведенные минимальные пары хорошо различаются аудиторами - представителями северожемайтских говоров. Циркумфлексы второго типа (~$_{(2)}$) аудиторы в подавляющем большинстве случаев воспринимают как интонацию, более близкую к циркумфлексу (восходящей, плавной интонации) литовского литературного языка; аудиторы, владеющие латышским языком, считают вторичный жемайтский акут (^(2)) идентичным прерывистой (lauztā) интонации латышского литературного языка (с.-жем. bû·s 'будет(-ут)' = лат. bûs 'т.ж.'), а вторичный циркумфлекс - латышской длительной (stieptā) интонации.

В диахроническом плане особенно интересным представляется первый (а) случай, так как он свидетельствует о возможной генетической связи вторичных типов интонаций с подвижными ("окситоническими") акцентуационными парадигмами прабалтийского языка (см. ниже, § 3).

I.2. Несколько позже сходное явление обнаружено и в восточно-литовских "утянских" и "паневежских" говорах (см.,напр., /2/, /3/). Аудитивные эксперименты, выполненные с представителями северных паневежских говоров, свидетельствуют о весьма четком противопоставлении: аудиторы правильно распознали до 84,5% предлагаемых минимальных пар типа płåũ.ke$_{(1)}$ 'плыл(а)' (инф. płåũ.ktь) : płåũ·ke$_{(2)}$ 'плавал(а)' (инф. płaukí·t̂ь) и łá·uže$_{(1)}$ 'ломал(а)' (инф. łá·uš̂t̂ь) : ła.úže$_{(2)}$ 'разламывал(а)' (инф. ła.úžệtь). Особенно хорошо различаются первичный и вторичный циркумфлексы – для некоторых пар получено до 94,3% правильных идентификаций.

Первичный ("баритонический") восточно-аукштайтский акут аудиторы почти единодушно оценивают как более резко падающую интонацию, а вторичный ("окситонический") циркумфлекс – как резко восходящую интонацию. По-видимому, это объясняется прежде всего различным соотношением длительности (и акустической энергии) компонентов дифтонга. По нашим предварительным обследованиям в паневежских говорах первый компонент оказался более длительным под первичным акутом (ср.: $\bar{x}_I$=I89 мс и $\bar{x}_2$ =I75 мс; $t=2,25 > t_{0,05}=I,97$), а вторичный циркумфлекс несколько увеличивает длительность второго компонента (ср.: $\bar{x}_2$ =I66 мс и $\bar{x}_I$=I50 мс; $t=3,48 > t_{0,001}=3,34$) /3/.

2.0. Более подробный анализ акустических признаков вторичных типов слоговых интонаций проводился на ЭВМ типа ЕС-I060.02 по программе "EGLE", составленной нами на языке программирования PL/1. В машину (как первичные данные) вводились результаты ручного измерения осциллограмм (измерялись амплитуда и длительность отрезков вокалического сегмента в 2-3 квазипериода). Вся дальнейшая аналитическая работа выполнена на ЭВМ, начиная с трансформации первичных измерений в физические единицы (децибеллы, герцы и т. д.) и кончая построением "точечных" усредненных графиков движения основного тона и интенсивности (причем на листингах также печатались точные средние значения избранных точек графика и их доверительные 95-процентные интервалы). Автоматически определялось расстояние от начала вокалического сегмента до пика интенсивности и основного тона, точки глоттализации, диапазоны и крутизна "восхождения" и падения интенсивности и тона и т.д. Кроме обычных статистических параметров (средних арифметических, стандартных отклонений, доверительных интервалов и др. /4/), машина также определяла коэффициенты корреляции таких явлений, как основной тон и интенсивность, основной тон и длительность, интенсивность и длительность.

Предварительная экспериментальная работа проводилась в Лаборатории экспериментальной фонетики, математическая обработка данных – в Вычислительном центре коллективного пользования ВГУ (консультант – доц. В. Ундзенас).

2.I.I. В северожемайтских говорах (исследовались глагольные формы типа буд. в. ka.îš(2) 'будет (-ут) чистить', rã.u.s(2) 'будет (-ут) рыть', kuôš(2) 'будет (-ут) цедить' и наст. в. kâ·iš(1) 'чистит(-ят)', rã·us(1) 'роет (-ют)', kûoš(1) 'цедит(-ят)' и др.), наиболее четко различаются первичный и вторичный акуты, реализуемые как прерывистая интонация. В тех случаях, когда слогоносителем является сложный дифтонг (/au/, /ai/ и др.), вторичный акут отличается от первичного длительностью слогоносителя ($\bar{x}_2=365\pm35$ мс, $\bar{x}_I=400\pm32$ мс, $t=4,I8 > t_{0,05}=2,00$), разницей основного тона и интенсивности первой и второй половины дифтонга (соответственно $\bar{x}_2$=6,I пт и $\bar{x}_I$=5,2 пт, $\bar{x}_2$=2 дБ и $\bar{x}_I$ = 2,6 дБ), различным относительным "расстоянием" до точки глоттализации ($\bar{x}_2=52\pm I4$%, $\bar{x}_I = 58 \pm I6$%, $t_{/\bar{A}/}=2,03 > t_{0,05}=2,00$), крутизной падения интенсивности и тона (соответственно $\bar{x}_2=I00\pm32$ дБ/с; $\bar{x}_I = 82\pm33$ дБ/с, $t=2,25 > t_{0,05}=2,00$; $\bar{x}_2=93\pm35$ пт/с, $\bar{x}_I=78\pm27$ пт/с, $t=2,95 > t_{0,05}=2,04$).

Практически во всех случаях более четкими оказались признаки интенсивности (см. рис. I) – основной тон выполняет как бы вспомогательную роль.

Резкий подъем интенсивности к концу дифтонга – носителя вторичного акута, по-видимому, и производит впечатление сильной глоттализации ("прерыва"), отмечаемой многими аудиторами. На кривых основного тона (см. рис. 2) это явление не наблюдается: все тоновые отличия сконцентрированы в начальном отрезке слогоносителя.

Результаты анализа акутированных слит-

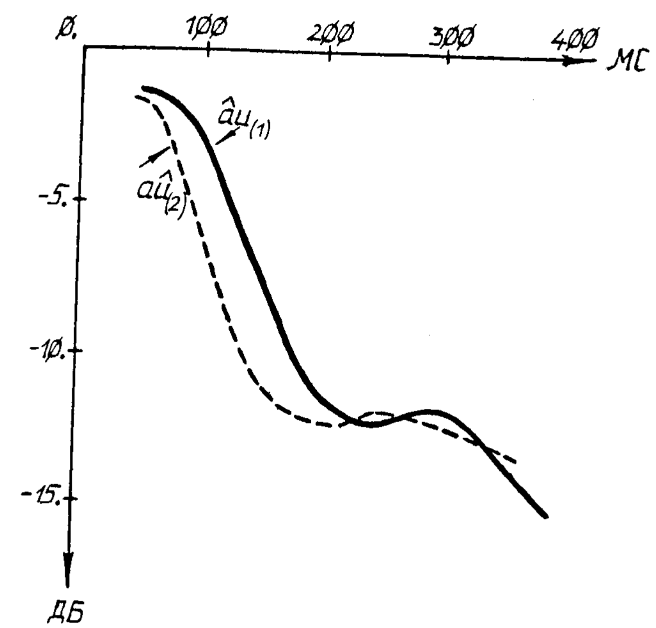


Рис. I. Обобщенные кривые движения интенсивности сложных акутированных дифтонгов северожемайтских говоров.

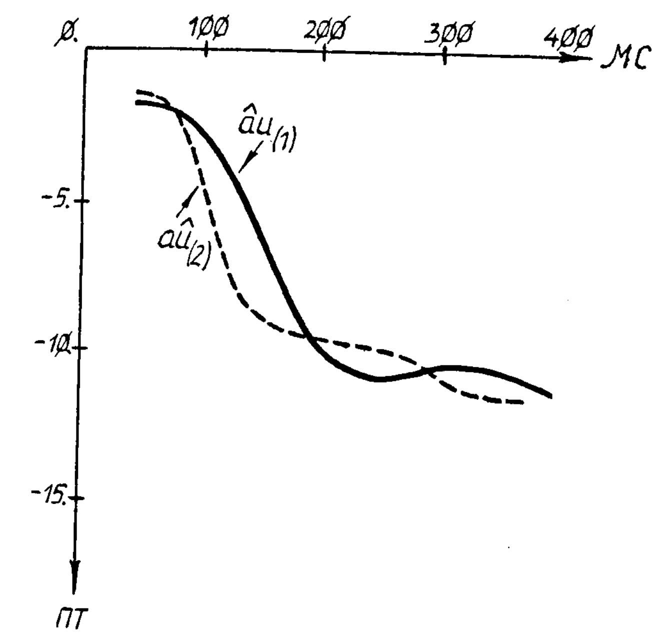


Рис. 2. Обобщенные кривые движения основного тона сложных акутированных дифтонгов северожемайтских говоров.

ных дифтонгов (/ie/, /uo/) свидетельствуют о тех же тенденциях. И в данном случае вторичные типы акута различаются длительностью ($\bar{x}_2=27I\pm38$ мс, $\bar{x}_I=3I2\pm4I$ мс, $t = 2,74 > t_{0,05}=2,06$), соотношением интенсивности первой и второй половины слогоносителя ($\bar{x}_2=6\pm3$ дБ, $\bar{x}_I=8\pm2$ дБ, $t=2,70 > t_{0,05}=2,06$), относительным "расстоянием" до точки глоттализации ($\bar{x}_2=53\pm I4$%, $\bar{x}_I=68\pm I8$%, $t=2,54 > t_{0,05}=2,06$), а также крутизной падения интенсивности ($\bar{x}_2=II4\pm32$ дБ/с, $\bar{x}_I = 83\pm22$ дБ/с, $t=2,93 > t_{0,05}=2,06$). Основной тон голоса оказался совсем незначимым – исследуемые просодемы, по всей вероятности, характеризуются лишь динамическими признаками.

2.I.2. Вторичные типы северожемайтского циркумфлекса (если судить по результатам нашего эксперимента) различаются значительно слабее. Обнаружены лишь следующие статистически значимые различия: длительность слогоносителя ($\bar{x}_2=38I\pm37$ мс, $\bar{x}_I=43I\pm39$ мс, $t=3,43 > t_{0,05}=2,06$), диапазон изменения ("восхождения") интенсивности ($\bar{x}_2=4\pm I$ дБ, $\bar{x}_I=2\pm I$ дБ, $t=3,00 > t_{0,05}=2,06$) и крутизна "восхождения" интенсивности ($\bar{x}_2 \approx 4I$ дБ/с, $\bar{x}_I = 2I$ дБ/с, $t=2,2I > t_{0,05} = 2,06$).

2.2.I. В восточно-аукштайтских (паневежских) говорах (исследовались глаголы одноразового и многократного действия brå̃u.ke(1) 'перечеркнул(а)' (инф. brå̃u.-kt̂ь), trá·uke(1) 'тянул(а)' (инф. trá·u-kt̂ь), brå̃u·ke(2) 'перечеркивал(а)' (инф. brauki·tь), tra.úke(2) 'дергал(а)' (инф. tra.úkęt̂ь) и др.), как и предполагалось по результатам предварительного аудирования, лучше различаются вторичные типы циркумфлекса.

Особенно четко различается среднее значение основного тона ($\bar{x}_2=-2,I\pm0,9$ пт, $\bar{x}_I=-2,5\pm0,8$ пт, $t=2,29 > t_{0,05}=I,98$), средний тон второй половины дифтонга ($\bar{x}_2=-2,3\pm0,8$ пт, $\bar{x}_I=-2,7\pm I,I$ пт, $t=2,44 > t_{0,05}=I,98$), минимальное значение тона ($\bar{x}_2=-4,I\pm I,9$пт, $\bar{x}_I=-5,8\pm3,2$ пт, $t=3,20 > t_{0,05}=I,98$), относительное "расстояние" до минимума тона ($\bar{x}_2=67,2\pm28,4$%, $\bar{x}_I=80,8\pm23,8$%, $t = 2,32 > t_{0,05}=I,98$), наконец – различный диапазон падения тона ($\bar{x}_2=3,9\pm2,5$ пт, $\bar{x}_I = 5,3\pm I,3$ пт, $t=2,I4 > t_{0,05}=I,98$).

Весьма значима и общая длительность

слогоносителя, только в данном случае (в отличие от жемайтских говоров) более длительным оказался вторичный циркумфлекс($\bar{x}_2$ =296±53 мс, $\bar{x}_I$=280±43 мс, $t_{/\bar{x}/}$=2,98>$t_{0,05}$ =2,0I).

Обобщенные кривые интенсивности получились почти изоморфными, а кривые основного тона свидетельствуют о весьма существенном различии (см. рис. 3).

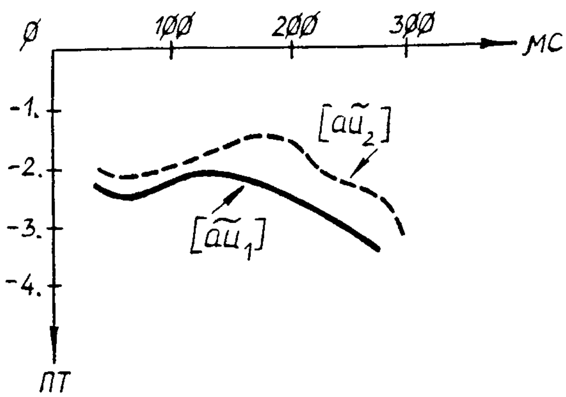


Рис. 3. Обобщенные кривые движения основного тона сложных циркумфлектированных восточно-аукштайтских дифтонгов.

2.2.2. В акутированных слогах восточно-аукштайтских (паневежских) говоров установлен лишь один статистически значимый признак - различное "расстояние" минимума интенсивности от начала дифтонга ($\bar{x}_2$=88±21%, $\bar{x}_I$=95±10%, t=2,15>$t_{0,05}$=1,99); о сходной тенденции свидетельствует и положение минимума основного тона ($\bar{x}_2$=72±24%, $\bar{x}_I$=82±22%, t=1,96>$t_{0,1}$=1,66).

3. Итак, в литовских говорах противопоставляются не только первичные типы слоговых интонаций (акут и циркумфлекс), но и вторичные их типы (первичный и вторичный акут, первичный и вторичный циркумфлекс). Точный фонологический статус вторичных интонаций пока не представляется вполне ясным: возможно, что их оппозиции сводятся к противопоставлению двух различных типов словесного ударения (существование которых в литовских диалектах уже доказано), хотя против такой трактовки можно выдвинуть и некоторые возражения.

В диахроническом плане наиболее важен вопрос о связи вторичных интонаций с прабалтийскими (и праиндоевропейскими) акцентуационными парадигмами и, более конкретно, с перемещениями ударения в различных формах, относящихся к одной и той же акцентуационной парадигме. Возникает серьезная дилемма, что считать первичным: "окситоническое" ударение словоформ, относящееся к балтийской подвижной ("окситонической") парадигме, или же вторичные слоговые интонации. Вторичные интонации можно объяснять как результат ретракции ударения с конечных слогов (ср. возникновение среднелатышской прерывистой интонации и сходные явления в литовских "баритонических" говорах), но принципиально возможна и противоположная точка зрения: ударение могло оказаться на конечном слоге известных словоформ в результате его перемещения с основ, обладавших прототипами современных "вторичных" интонаций (ср. сходное более позднее явление, описываемое законом Фортунатова - де Соссюра).

Мы склоняемся к первой альтернативе, хотя доказать ее "единственность", по-видимому, пока нет реальной возможности.

ЛИТЕРАТУРА

1. Гирденис А. Опыт морфонологической интерпретации северожемайтской аттракции ударения // Baltistica. 1980. Т.18(2). С. 179-188.
2. Качюшкене Г. Й. Фонологическая система северопаневежского диалекта литовского языка: (Просодия и вокализм): Дис. ... канд. филол. наук. Вильнюс, 1984.
3. Kačiuškienė G. Antriniai priegaidžių tipai šiaurinių panevėžiškių tarmėje// Mokomojo ir auklėjamojo proceso organizavimas: Pranešimų tezės. Šiauliai, 1983. P. 179-180.
4. Урбах В. Ю. Статистический анализ в биологических и медицинских исследованиях. М.: Медицина, 1975.

ДИНАМИКА ОППОЗИЦИЙ СЛОГОВЫХ ИНТОНАЦИЙ В ФОНОЛОГИЧЕСКИХ СИСТЕМАХ ДИАЛЕКТНОЙ И ГОРОДСКОЙ РЕЧИ /НА МАТЕРИАЛЕ ЛИТОВСКОГО ЯЗЫКА/

ЛАЙМА ГРУМАДЕНЕ

Институт литовского языка и литературы АН ЛССР
Вильнюс, ЛитССР, СССР 232043

БОНИФАЦАС СТУНДЖЯ

Вильнюсский госуниверситет
Кафедра литовского языка
Вильнюс, ЛитССР, СССР 232017

РЕЗЮМЕ

В докладе приводятся экспериментальные данные, свидетельствующие об относительно малой значимости просодических характеристик (по-видимому, за исключением признака длительности) слоговых интонаций дифтонгов не только в литовской городской речи, но и – в меньшей степени – в восточном диалекте.

В настоящее время можно считать доказанным, что слоговые интонации северо-западных (жемайтских) диалектов литовского языка носят преимущественно динамический, а не музыкальный характер /1/. Природа слоговых интонаций восточных (и южных) диалектов является более проблематичной. Особенно важным (и, кроме того, весьма дискуссионным) до сих пор остается вопрос об относительной нивеляции данной просодической оппозиции на восточно-литовских монофтонгах /2/. Существенные изменения акцентуационной системы, наблюдаемые в городской речи /3/, указывают на определенные существенные сдвиги в системе слоговых интонаций.

В данном исследовании сопоставляются просодические характеристики слоговых интонаций в речи представителей восточных аукштайтов и уроженцев г.Вильнюс. Для инструментального анализа были подобраны слова с дифтонгами ai, au, ei, слоговые интонации которых вполне четко различаются как во всех диалектах, так и в городской речи. Дифтонги ai, au, ei были реализованы между глухими согласными или после паузы: táiko – užkaĩto, táiko – paĩko, káišo – paĩšo, táuko – taũpo, šáuk – šaũk, káukė – kaũkė, káušo – kaũšo, áuk – aũk, kéikia – peĩkia, kéikė – peĩkė, kéik – peĩk. Данные слова "вставлялись" в краткие предложения, в которых они занимали среднюю позицию, и были зачитаны с восходящей (↑) и нисходящей (↓) фразовыми интонациями без логического ударения на исследуемых словах, а также в связном тексте (↑↓) под неэмфатическим фразовым ударением. Дикторами были два представителя восточно-аукштайтского утянского диалекта и два уроженца г.Вильнюс, не владеющие ни одним традиционным литовским диалектом.

Первичные данные визуально измеренных 264 осциллограмм были подвергнуты математико-статистическому анализу в Вычислительном центре коллективного пользования ВГУ по программе "EGLE" (язык программирования PL/1; составил А.Гирденис), которая автоматически определяет и статистически обрабатывает все основные просодичесие характеристики интонаций (см./4/).

Результаты анализа свидетельствуют о том, что самым постоянным просодическим индикатором слоговых интонаций дифтонгов исследуемых восточно-литовских говоров и речи вильнюсцев является длительность (акутированные дифтонги по общей длительности явно превосходит соответствующие циркумфлектированные дифтонги (на что ещё в 1925 г. указал Р.Экблум /5/, ср. также /6/)). Данное различие более ощутимо в городской речи: если у представителей диалекта доверительные интервалы средних значений общей длительности éi, éu, éi : ei͂, eu͂, ei͂ не пересекаются (в тексте – и значительно), то у представителей городской речи этого не наблюдается ни в одной позиции (см. таблицу).

Таблица

Общая длительность сложных дифтонгов éi, éu, éi : ei͂, eu͂, ei͂

Интонация	n	$\bar{X}$	±95%	$t \gtrless t_{05}$
		диалект		
↑	22 22	207.5 174.7	193.3÷221.5 149.8÷199.6	1.97 < 2.02
↓	21 23	199.4 169.0	179.3÷219.6 146.3÷191.8	2.06 > 2.02
↑↓	25 24	182.7 168.6	160.3÷205.2 146.6÷190.6	0.93 < 2.01
		город		
↑	20 21	174.4 139.2	165.4÷183.5 129.5÷148.9	5.55 > 2.02
↓	21 21	176.7 142.3	167.4÷186.0 133.6÷150.9	5.64 > 2.02
↑↓	21 22	176.9 141.2	165.3÷188.5 130.9÷151.4	4.82 > 2.02

Поскольку акут в восточных говорах и городской речи является маркированным членом бинарной просодической оппозиции, бóльшую длительность акутированных дифтонгов следует считать вполне закономерной (в западных диалектах, где маркированным членом является циркумфлекс, имеет место обратное соотношение /7/; в новейших исследованиях утверждается, что в литературном языке длительность дифтонгов – инвариантный признак слоговых интонаций /8/).
Фонетической предпосылкой бóльшей длительности акутированных дифтонгов несомненно следует считать четкую реализацию и заметное удлинение первого компонента. Такое удлинение, свойственное восточным и южным диалектам, охватывает также и дифтонги типа íe, úo(>í.e, ú.o/ í·e, ú·o), сохраняющие "краткость" (или, точнее, ненапряженность) i, u в нормативной литературной речи /9/.
Релевантность других просодических признаков – интенсивности, основного тона – в большинстве случаев зависит от конкретной позиции.
Что касается представителей диалекта, то максимальное число релевантных различий наблюдается в тексте (компенсация отсутствия значимости различия акутированных и циркумфлектированных дифтонгов по длительности?): акутированные дифтонги отличаются меньшим расстоянием между началом дифтонга и максимумом интенсивности (30.4%), чем циркумфлектированные (46.6%), а также более узким диапазоном восхождения и более низкой второй частью кривой основного тона. Циркумфлексу свойственна (особенно в позиции с нисходящей фразовой интонацией) бóльшая разница между I и 2 частями интенсивности. Для акутированных слогов в диалекте характерны сильные корреляции между средней длительностью и средней интенсивностью или тоном, между средней интенсивностью и средним тоном (за исключением позиции с нисходящей фразовой интонацией), тогда как для циркумфлектированных – лишь последняя корреляция.
В речи вильнюсцев (кроме значимости общей длительности) следует указать на сильную корреляцию между интенсивностью и основным тоном; она обнаруживается в основном только в циркумфлектированных слогах. Например, при нисходящей фразовой интонации она является очень сильной (Z = 0.3682 > > 0.1600), а в предложениях с восходящей фразовой интонацией отсутствует.
Сравнение просодических свойств речи представителей диалекта и города показывает, что дифтонги ai, au, ei в вильнюсской речи являются более краткими, чем соответствующие дифтонги восточно-аукштайтского диалекта (разница статистически значима, за исключением акутированных дифтонгов в тексте (конкретные данные ср. по таблице)).
Максимальное число релевантных различий между диалектной и городской речью наблюдается у акутированных дифтонгов, реализуемых с нисходящей фразовой интонацией. В этой позиции данные дифтонги восточно-аукштайтского диалекта отличаются многими характеристиками интенсивности (например, более низкой I частью кривой, меньшей разницей между I и 2 частями слогоносителя, бóльшим расстоянием до начала минимума и меньшим расстоянием между началом дифтонга и максимумом, более узким диапазоном восхождения и меньшей крутизной падения) и основного тона (например, бóльшим расстоянием между началом дифтонга и максимумом, меньшей крутизной восхождения, более узким диапазоном нисхождения и др.). Корреляция интенсивности и основного тона как в диалектной, так и в городской речи слабая. Многие различия, по-видимому, объясняются воздействием фактора длительности.
Циркумфлектированные дифтонги под нисходящей интонацией менее различны. Кроме того, и в диалекте, и в городской речи наблюдается сильная корреляция между интенсивностью и основным тоном. Циркумфлексу восточно-аукштайтского диалекта характерно бóльшее расстояние до начала минимума интенсивности и основного тона.
Примерно те же различия в просодических особенностях циркумфлектированных дифтонгов диалектной и городской речи наблюдаются и с восходящей интонацией. Только в данном случае для диалекта характерен более широкий диапазон падения интенсивности и основного тона, а также бóльшая крутизна падения основного тона.
Что касается акутированных дифтонгов в позиции с восходящей интонацией, то основные различия характеристик основного тона между диалектной и городской речью сохраняются, а различия характеристик интенсивности отсутствуют, за исключением расстояния до начала ми_нимума.

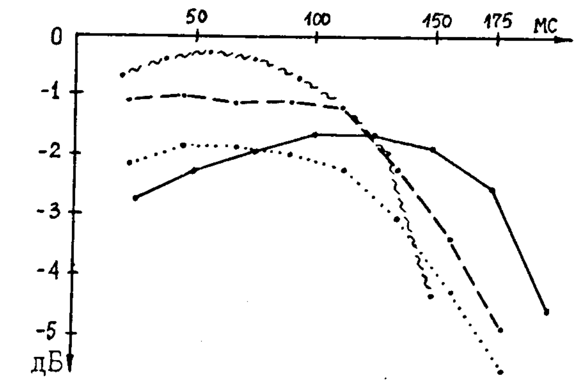


Рис.I. Обобщенные кривые движения интенсивности сложных дифтонгов с нисходящей фразовой интонацией.

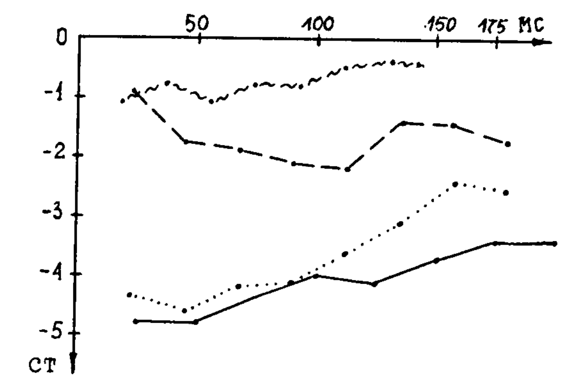


Рис.2. Обобщенные кривые движения основного тона сложных дифтонгов с нисходящей фразовой интонацией.

——— – акут } вост. аукштайты
– – – – циркумфлекс }
······ – акут } вильнюсцы
~~~ – циркумфлекс }

Математико-статистический анализ осциллографических данных свидетельствует о том, что в диалектной системе слоговых интонаций дифтонгов сохранилось бóльшее число релевантных просодических характеристик, чем в городской речи. Постоянным общим признаком обеих систем является длительность, по всей вероятности, вытекающая из различного качества первого компонента акутированных и циркумфлектированных диф-
~~~

тонгов. Следовательно, на востоке литовского языкового ареала просодические различительные признаки слоговых интонаций постепенно теряют самостоятельную значимость даже на дифтонгах, причем в городской речи этот процесс происходит более энергично и последовательно. Это явление неизбежно должно привести к ослаблению и даже исчезновению оппозиции слоговых интонаций на монофтонгах, где просодические признаки не компенсируются качественными. По-видимому, этим можно объяснить результаты прелиминарного аудиотестирования, указывающие на то, что в восточном литовском (в частности утянском) говоре слоговые интонации монофтонгов (и ie, uo) лучше различаются в тех случаях, когда слогоноситель, например, o и uo, может легко менять качество (ɔ́/ °ã, úɔ/ᵘã).

ЛИТЕРАТУРА

/1/ Cirdenis A. Prozodinės priegaidžių ypatybės šiaurės žemaičių tarmėje//Eksperimentinė ir praktinė fonetika. Vilnius, 1974. P.193-194.

/2/ См. Zinkevičius Z. Lietuvių dialektologija. Vilnius, 1966. P.33 (и литературу);Kosienė O. Rytų aukštaičių uteniškių monoftongų priegaidės//Kalbotyra. 1982. T.33(1); Garšva K. Svarbesnės šiaurės vakarų panevėžiškių fonologijos ypatybės//Baltistica.1982. T.18(1); Kačiuškienė G. Kuo gali skirtis šiaurinių panevėžiškių priegaidės//Kalbotyra.1985. T.36(1); Girdenis A., Pupkis A. Pietinių vakarų aukštaičių priegaidės (prozodiniai požymiai)//Eksperimentinė ir praktinė fonetika. Vilnius, 1974. P.116-119.

/3/ Grumadienė L. Sociolingvistinis vilniečių lietuvių kalbos tyrimas : konsonantizmas ir akcentuacija//Lietuvių kalbotyros klausimai. 1987.T.27.

/4/ Гирденис А., Качюшкене Г. Вторичные типы слоговых интонаций в литовских диалектах//Материалы XI международного конгресса фонетистов. Таллин, 1987.

/5/ Ekblom R. Ouantität und Intonation im Zentralen Hochlitauischen. Uppsala, 1925. S. 38, 64.

/6/ Anusienė L. Kirčiuotų akūtinių ir cirkumfleksinių dvibalsių trukmė lietuvių bendrinės kalbos frazėse//Kalbotyra.1984. T.35(1). P.14.

/7/ Girdenis A. Op. cit.P.175.

/8/ Pakerys A. Lietuvių bendrinės kalbos prozodija. - Vilnius, 1982. P.156.

/9/ Pakerys A. Lietuvių bendrinės kalbos fonetika. - Vilnius, 1986. P.317.

DURATION OF LONG STRESSED VOWELS IN PRESENT-DAY LITHUANIAN UTTERANCES

LILIJA ANUSIENĖ

Dept. of Foreign Languages
Vilnius Civil Engineering Institute
Vilnius, Lithuania, USSR 232040

ABSTRACT

This paper reports on the results of an investigation into the duration of acute and circumflex vowels and vocalic diphthongs in extended speech contexts. The results obtained lead us to the conclusion that duration is not the main factor in the differentiation of accent type.

In Lithuanian, long monophthongs, and vocalic or mixed diphthongs (vowel plus either liquid or nasal consonant) in stressed position may have falling or rising accent. The terms "falling" and "rising" retain musical connotations, and the actual nature of the intonation is in doubt, so for practical purposes it is better to use the neutral terms "acute" and "circumflex" /I/. These terms refer only to the names of the signs used to mark the accent /2/. Some investigators have proposed that the most notable prosodic feature of Lithuanian accent is vowel duration /3/, with the circumflex vowels being longer than acute ones /4/.

The duration of syllabic neuclei in the Lithuanian colloquial language has been experimentaly investigated by many linguists. Previous studies used the following as test material: 1) isolated two-syllable words /5/, 2) minimal pairs in isolation and 3) minimal pairs placed at the beginning, in the middle, or at the end of short phrases /6/.

The purpose of this research was to study the duration of acute and circumflex vowels and vocalic diphthongs. The experimental data consisted of 128 utterances, typical of the Standard Lithuanian language, recorded on magnetic tape by 3 male and 2 female subjects. Measurements were obtained from intonograms.

In the experimental material the vowels and diphthongs under investigation are found in various phonetic environments and in various positions in the phrase. The vowels in question are found in all possible positions in the word and in the phrase. So as to compensate for the influence of the position of the word in the phrase the experimental phrases were constructed so that the vowel is found an equal number of times in each position. In order to compensate for differences in absolute duration in different positions computations are based on relative differences in duration. The data for each subject were individually analyzed, but since the same corpus was used for each subject we can also contrast the data on vowel and diphthong duration for all the subjects as a group.

Previous studies which examined the duration of long stressed vowels revealed that: 1) in isolated two syllable words the duration of circumflex vowels is always greater than that of acute vowels /7/, 2) in short phrases material results indicated that duration of circumflex vowels is greater in 86.7% of the cases /8/.

The results of the present research revealed that there is almost no difference in the duration of circumflex and acute vowels in extended speech contexts.

Lehiste /9/ claims that in the range of the durations of speech sounds - usually from 30 to about 300 msec - the just-noticeable differences in duration are between 10 and 40 msec. In our experimental material the difference in duration of vowels with different accent types is less than 10 msec in 53.3% of the cases, is less than 20 msec in 40% of the cases, and slightly exceeds 20 msec in 6.7% of the cases. In subject 3 there is a substantial difference in duration between acute and circumflex /i/ (1:1.23) but there is almost no difference for subjects 1 and 4 (1:1.02; 1:1.01). The two /u/'s with different accent types do not differ for subjects 3 and 4 (1:1.05; 1:1.04), but do differ substantially for subject 1 (1:1.2). There are no other cases of clear differences in duration between acute and circumflex vowels.

The results obtained revealed that there is no significant difference in duration of the acute and circumflex vowels. This leads us to the conclusion that duration is not the main factor in the differentiation of accent type in the Lithuanian colloquial language.

Some Lithuanian linguists /10/ claim that circumflex diphthongs are longer than acute ones. Pakerys, Plakunova and Urbelienė /11/ claim that the diphthongs /au/, /ai/, /ei/ are almost equal in duration, irrespective of the type of accent. The results of the present research revealed that the acute diphthongs are longer than the circumflex ones. Substantial difference in duration between different accent types is characteristic of the pronunciation of all the subjects in the case of the diphthong /ai/, for three subjects in the case of /au/ and for one subject in the case of /ei/. As only one case out of fifteen (i.e. 6.67% of the cases) shows substantial difference in the duration of the diphthongs /ei/, /ie/, /uo/ in favour of

the acute accent type, we may claim that there is essentially no difference in the duration of the above mentioned circumflex and acute diphthongs. It should be noted that the range of duration of circumflex diphthongs in female subjects is very small, so that in their speech there is no significant difference in duration for all the circumflex diphthongs under investigation.

The given data allow us to arrange the diphthongs, irrespective of the accent type, in order of decreasing duration: au → ai → ei → uo → ie.

From the above evidence we may conclude that:

1. There is no significant difference in duration between acute and circumflex vowels.
2. There is a substantial difference in the duration of the diphthongs /ai/ and /au/ pronounced with different accent type.
3. Where (2) is relevant, acute diphthongs are longer than circumflex ones.
4. There is no significant difference in duration for the diphthongs /ei/, /uo/, /ie/ pronounced with different accent type.
5. According to their duration, diphthongs may be classified into three groups, irrespective of accent type: /au/, /ai/ - the longest, /ei/ - medium, /uo/, /ie/ - the shortest.
6. Circumflex diphthongs in the speech of female subjects show no difference in duration.

REFERENCES

/1/ A.Girdenis, Fonologija, Vilnius, 1981. 188.

/2/ Z.Zinkevičius, Lietuvių kalbos istorinė gramatika I, Vilnius, 1980. 45.

/3/ A.Girdenis, A.Pupkis, "Pietinių vakarų aukštaičių priegaidės (prozodiniai požymiai)", Eksperimentinė ir praktinė fonetika (Eksperimentinės fonetikos ir kalbos psichologijos kolokviumo medžiaga VI), Vilnius, 1974. 116.

/4/ A.Girdenis, A.Pupkis, Op. cit., 1974. 116; A.Pakerys, T.Plakunova, J.Urbelienė, "Otnositel'naja dlitel'nost' glasnyx litovskogo jazyka", Kalbos garsai ir intonacija (Eksperimentinės fonetikos ir kalbos psichologijos kolokviumo medžiaga IV), (Lith. resume 54), Vilnius, 1970. 45.

/5/ A.Pakerys, T.Plakunova, J.Urbelienė, Op. cit., 1970. 30-54.

/6/ A.Girdenis, A.Pupkis, Op. cit., 1974. 107-125.

/7/ A.Pakerys, T.Plakunova, J.Urbelienė, Op. cit., 1970. 45.

/8/ A.Girdenis, A.Pupkis, Op. cit., 1974. 111.

/9/ I.Lehiste, Suprasegmentals, MIT Press, Cambridge, 1970. 13.

/10/ V.Vaitkevičiūtė, "Lietuvių kalbos

balsių ir dvibalsių ilgumas arba kie-kybė", Lietuvių kalbotyros klausimai III, Vilnius, 1960. 217; G.Daugirdaitė, "Dabartinės lietuvių literatūrinės kalbos sutaptinių dvibalsių /ie/ ir /uo/ trukmė", Kalbos garsai ir intonacija (Eksperimentinės fonetikos ir kalbos psichologijos kolokviumo medžiaga IV), Vilnius, 1970. 69.

/11/ A.Pakerys, "Lietuvių literatūrinės kalbos sudėtinių dvibalsių au, ai, ei akustiniai požymiai", Eksperimentinės fonetikos ir kalbos psichologijos kolokviumo medžiaga III, Vilnius, 1968. 106-7; A.Pakerys, T.Plakunova, J.Urbelienė, "Otnositel'naja dlitel'nost' diftongov litovskogo jazyka", Garsai, priegaidė, intonacija (Eksperimentinės fonetikos ir kalbos psichologijos kolokviumo medžiaga V), (Lith. resume 36), Vilnius, 1972. 7.

PITCH ACCENTS IN STANDARD LITHUANIAN

VALERIJA VAITKEVIČIŪTĖ

State Conservatoire
Vilnius, Lithuania, USSR 232000

ABSTRACT

In Standard Lithuanian there is no overall scheme for the realisation of pitch accents. A common feature in speakers of the Standard language and also in those of various dialects is the constant presence of opposition of acute and circumflex accents, while the choice of phonetic characteristics used in opposition and the way they were used varied from dialect to dialect. Prosodic distinctions are found in the difference in level of amplitude and fundamental frequency and not in their contours.

INTRODUCTION

Contemporary researchers into Standard Lithuanian pitch accents /1,2,3/ have attempted to find an overall scheme for their realisation in the parameters of difference in fundamental frequency, in intensity and in duration. They took averages of data received from speakers of both sexes (with varying diaposon of voice) who also had different dialect origins. Their estimates of durational, fundamental frequency and amplitudinal difference in pitch accents of vowels were based only on the number of cases and did not take into account whether or not such differences were of any significanse for perception. Researchers have also failed to attach significance to the following remarks of the well-known dialectician Z.Zinkevičius /4/: "Speakers of dialects who subsequently learn Standard Lithuanian pronounce monophthongs and diphthongs with the pitch accents of that dialect. They do not acquire the pitch accents of Standard Lithuanian, that is of the language spoken in the southern part of Western Aukštaičiai."

Pitch accents were studied from oscillographic recordings of normal and whispered speech and from listening tests using segmented quasi-homonyms as stimuli. The following parameters were investigated: duration, amplitude, fundamental frequency proportional energy of stressed vowels (the amount of total energy per msec), total energy of unstressed vowels as well as pitch fluctuation in adjacent syllables.

The work presents data from a recent oscillographic study of pitch accents in isolated disyllabic quasi-homonyms of speakers of Standard Lithuanian from different regions that has shown all the dialects to have a continual opposition of accent, while choice of phonetic characteristics and the manner of their use varied from dialect to dialect. Two speakers spoke the Kapsai dialect: Sp.1, 4; the Veliuoniškiai dialect was represented by Sp.2; the Dzūkai dialect by Sp.5. Amplitude, fundamental frequency and proportional energy were measured for vowels as a whole and for vowel parts (I, II, III): of the first and second components of diphthongs and monophthongs. This method was used to gather information concerning amplitudinal, fundamental frequency and proportional energy difference in different pitch accents of vowels in identical parts of the vowels. This identified the part which carries information about differences between the pitch accents in each of the various parameters. Analysis of the vowel part by part makes it possible to define the difference between pitch accents occuring, not in the contours of amplitude and fundamental frequency, but in the uneven level of these parameters as a whole. Only in this way is it possible to identify the particular part of the vowel where compensation of one parameter another takes place, to find out where correlation between them occurs and to find out which parameter is most important.

Differences of pitch accents in duration, amplitude and fundamental frequency were expressed in per cent and compared in pairs by the sign criterion (sign test) P = 0.05. First and foremost, we estimated all the differences revealing this tendency, disregarding their contribution to perception. The significance of differences in duration and amplitude as postulated by Weber and Fechner, and the significance of differences in fundamental frequency (tone) as postulated by Flanagan and Saslaw. Only these perceptually significant differences were later taken into consideration. Data on duration differences also included a record of the differences in the type of vowels under study.

PITCH ACCENTS IN SPEAKERS OF KAPSAI ORIGIN (Sp.1 AND 4)

In kapsai dialect (Sp.1 and 4), the most important features were amplitude (especially for Sp.4) and duration. In the pronunciation of Sp.1, information on vowel differences in amplitude, depending of the type of pitch accent, was contained in 1) an entire vowel (monophthong or diphthong), 2)an entire monophthong or the first component of a diphthong, 3)the first and third parts of a vowel. Sp.1 showed differences in whole vowels in 82% of the cases, significant differences in 55% of the cases. For 2) we obtained 91% and 55%, respectively. For 3) 82% and 64% for the Ist part of a vowel and 91% and 73% for the IIIrd part of a vowel. In all the situations mentioned above, the stressed vowel with acute accent had greater intensity than the vowel with circumflex accent. Data values expressing the tendency shown in points 1),2),3) by the sign criterion (P = 0.05) were labelled "+", and the significant difference in amplitude was called "-".

Therefore, the amplitude in phonetic realization of pitch accents in the pronunciation of Sp.1 is highly important. On the one hand, it is distinguished by a stable level within the vowels and expresses the tendency. On the other hand, the small number of quasi-homonyms where differences in amplitude were significant, indicates a certain lack of independence of this parameter. The same may be said concerning the differences in duration of vowels with various pitch accents. Duration differences in vowels were expressed in 100% of the cases; differences were significant in 64% of the cases, while dif ferences significantly correlated with the character in only 55% of the cases. The expressed duration differences by the sign criterion (P = 0.05) were considered "+", and significant differences in all the previously mentioned cases were "-".

Comparison of data on the ratio of amplitude and duration leads to the following conclusion. In the pronunciation of Sp.1 the uneven level of amplitude within vowels with different pitch accents is supported by their difference in duration: the proportional energy of the whole stressed vowel with acute accent is greater than that of a corresponding vowel with circumflex accent in 91% of the cases. The Ist, the IIIrd parts in 82% of the cases. In all cases data values for the sign criterion (P = 0.05) were "+", indicating the correlation of these two parameters.

Difference in fundamental frequency of vowels with different pitch accents for Sp.1 in 64% of the cases were expressed and significant only in the Ist part of the vowels. In both cases data values for the sign criterion (P = 0.05) were "-". Factors that witness its participation were as follows. First, the shift of maximum amplitude to the first part of the vowel in whispered speech. Secondly, the lack of significant differences in these vowel parts in amplitude and in duration. Third to some extent the test values for perception: pitch accents in quasi-homonyms with deleted initial consonants and onglides of vowels in some pairs were recognized as different, while in others they lost information concerning differences and were taken for identical acute accents (Fig.1).

In the pronunciation of Sp.1, the difference in fundamental frequency between syllables acted as an auxiliary: the differences in fundamental frequency between the last part of the vowel with acute accent and a following unstressed vowel was greater than the corresponding difference between the vowel with circumflex accent and a following unstressed vowel in 73% of the cases.

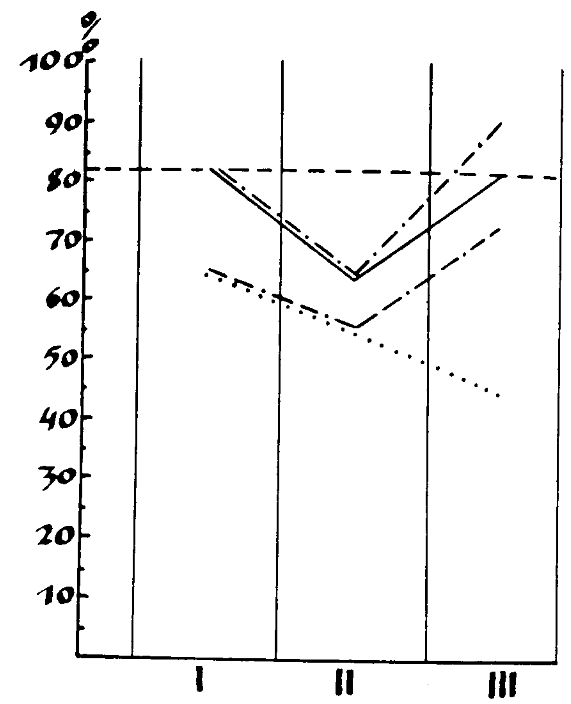


Fig.1. Speaker 1. Difference between vowels with acute accent and the corresponding vowels with circumflex accent. —·—· amplitude difference shown; —·—·—· significant amplitude difference; ——— proportional energy difference; ········ significant fundamental frequency difference; I, II, III corresponding parts of vowels; - - - - positive data value for P = 0.05.

In the pronunciation of Sp.4 duration differences of vowels were expressed in 100% of the cases, significantly so in 92%, and significantly corresponding to the character in 84% of the cases. Data values (P = 0.05) in all three cases were "+", indicating the independence of durational differences in vowels. Information on amplitude differences in pitch accents was carried by 1)the whole vowel, 2)the whole monophthong or only by the first component of the diphthong, 3)Ist and IInd parts of the vowels. Differences in whole vowels were marked in 75% of the cases, significantly so in 50% of the cases; in monophthongs and only the Ist components of diphthongs they were expressed in 75% of the cases, significantly in 58% of the cases, in the Ist parts in 84% of the cases, significantly in 50%; in IInd parts differences were expressed and significant in 84% of the cases. Data values expressing the tendency shown in points 1),2),3) by the sign criterion (P= 0.05) were "+". Data values of significant differences in amplitude in all the aforementioned points by the sign criterion (P= 0.05) were "-", excepting the IInd parts, where data values were "+". The fact that amplitude differences are of prime importance in the opposition of pitch accents was confirmed by the perception tests. In the pronunciation of Sp.4, the listeners could not discriminate even dynamically marked differences in the stressed syllable.

Difference in fundamental frequency of vowels with different pitch accents in 58% of the cases were expressed and significant in the IInd and IIIrd parts of the vowels (Fig.2).

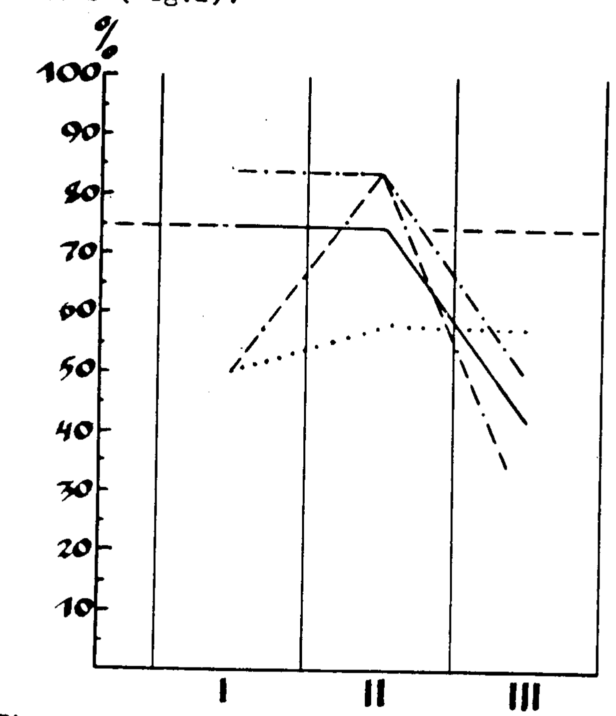


Fig.2. Speaker 4. Explanatory notes as at fig.1.

PITCH ACCENTS IN SPEAKER OF VELIUONIŠKIAI ORIGIN (Sp.2)

The most important features were duration and fundamental frequency. Duration differences of vowel sound were significantly marked and in keeping with the general character in 100% of the cases. Data values of significant difference according to the sign criterion (P= 0.05) were "+". Difference in fundamental frequency level occurred in all parts of the vowel, and greater amplitude and higher fundamental frequency were characteristic of certain parts of the vowel with acute accent in comparison with the vowel of circumflex accent. Differences in vowels with various pitch accents were marked and significantly so in 90% of the cases of whole vowels and monophthongs on the same grounds with the Ist components of diphthongs. In the Ist, IInd and IIIrd parts of vowels, differences were expressed and significantly so in 80% of the cases. Data values for the sign criterion (P= 0.05) were "+". Difference in amplitude of the whole vowel was expressed in 90% of the cases; differences in monophthongs in the same manner, with the Ist components of the diphthongs in 80% of the cases; of the first parts in 90% of the cases. Data values pressing tendency for the sign criterion (P= 0.05) were "+", significant differences in all the points were "-". In addition, the differences in vowels with acute and circumflex accents were supported by the differences expressing the tendency, of post-stressed syllables in total energy (in 80% of the cases), and also by the differences inclined toward tendency of the fundamental frequency between syllables (70%).

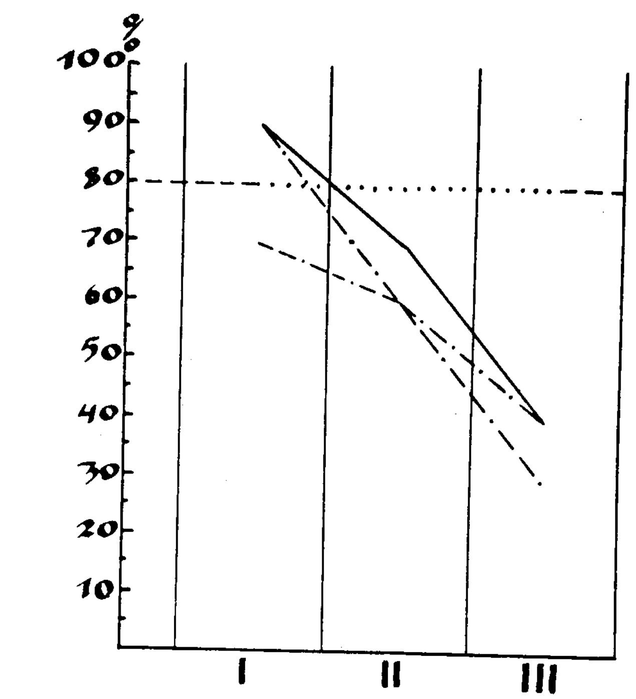


Fig.3. Speaker 2. Explanatory notes as at fig.1.

PITCH ACCENTS IN SPEAKER OF DZŪKAI ORIGIN (Sp.5)

In the speech of Sp.5, a representative of the dzūkai dialect, the important role in phonetic realization of vowels with different pitch accents was played by differences in duration (These were expressed in 90% of the cases and significantly corresponded to the character in 80% of the cases), by differences in fundamental frequency modulation between syllables (Differences in fundamental frequency between vowels with acute accent and post-stressed vowels in 80% of the cases were smaller than those between vowels with circumflex accent and post-stressed vowels), and by differences in total energy of post-stressed vowels (After acute accent the total energy was greater than after circumflex accent in 80% of the cases). Data values of significant differences in all aforementioned cases by the sign criterion (P= 0.05) were "+" (Fig.4).

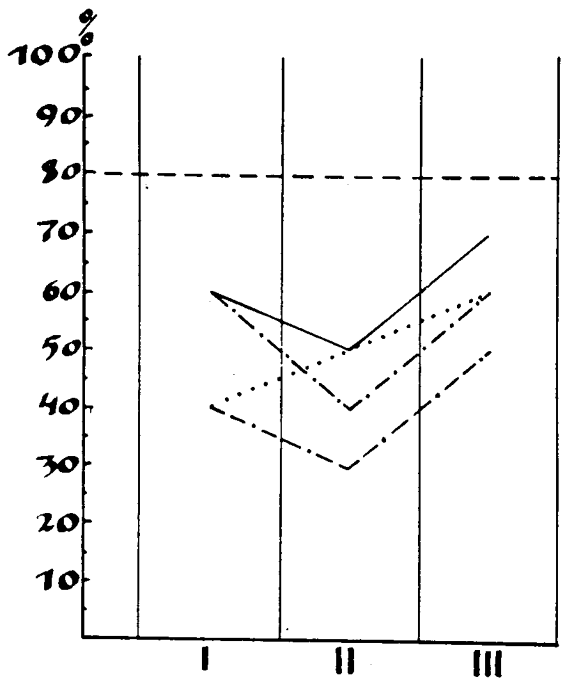


Fig.4. Speaker 5. Explanatory notes as at fig.1.

CONCLUSION

The data which we have obtained appears to show that the hitherto prevailing theory of the exixtence of an overall for the realisation of pitch accents of speakers of Standard Lithuanian irrespective of their original dialect is groundless. However, these investigations can at best, serve only as the starting point of a great deal of further work for those researchers investigating the prosody of Lithuanian, both in the standard language and in its dialects.

REFERENCES

/1/ A.Girdenis, A.Pupkis, "Pietinių vakarų aukštaičių priegaidės (prozodiniai požymiai)", Eksperimentinė ir praktinė fonetika V, Vilnius, 1974.
/2/ A.Pakerys, "Lietuvių bendrinės kalbos prozodija", Vilnius, 1982.
/3/ A.Pakerys, "Lietuvių bendrinės kalbos fonetika", Vilnius, 1986.
/4/ Z.Zinkevičius, "Lietuvių dialektologija", Vilnius, 1966.

TYPES OF SYLLABLE TONEME IN THE ZIEMERI VARIANT OF HIGH LATVIAN DIALECT

DACE MARKUS

Dept. of Languages and Literature
Latvian State University
Riga, Latvia, USSR 226000

ABSTRACT

The research permits to confirm the supposition concerning the functionating of two types of syllable toneme in the Eastern variants of the High Latvian dialect. The answer to the discutable question on the place of glottalization in the monophthongs and diphthongs with a broken (acute) syllable toneme is searched for.

INTRODUCTION

At present the amount of experimental phonetic research in the field of Latvia dialects leaves much to be desired. We have attempted to investigate a variant (sub-dialect) of a very peculiar and only partly explored High Latvian dialect called Augšzemnieku. The sub-dialect under consideration is still widely used basically in everyday life. The Ziemeri sub-dialect may be heard in the north-eastern part of Latvia adjacent to the border of the Estonian SSR. However, the Estonian language has not affected the Ziemeri sub-dialect to any notable extent. The variant in question belongs to the eastern group of High Latvian dialect. A distinctive feature of these sub-dialects is the monophthongization of the common Latvian diphthong ie (>ī) and uo (> ū), e.g. sìva <sieva 'wife'; ùla<uola 'egg'.

Two types of syllable tonemes function in the Ziemeri sub-dialect, namely, the so-called falling (\) and broken (glottalized; ∧), e.g. rèit 'to swallow' and reît 'tomorrow'. The level syllable toneme (~) occurring in the Latvian standard language is substituted by the falling syllable toneme in the Ziemeri sub-dialect, e.g. laĩme 'happiness', mãte 'mother', saũle 'sun', in the Ziemeri sub-dialect are pronounced as làima, mùotę or mùota, sàula. The two types of the syllable toneme were likewise distinguished by us in the sub-dialects used in the areas adjacent to that of the Ziemeri sub-dialect, namely, the sub-dialects of Alūksne, Jaunlaicene, Jaunroze, Karva and Veclaicene. To differentiate and identify a variety of syllable toneme, we investigated the fundamental pitch, intensity and the spectrum dynamics of vowels (vocalic centre of syllable).

So in the Ziemeri sub-dialect two types of syllable pitch are most strikingly correlated according to the dynamics (characteristic motion) and direction of the fundamental pitch in vowels. They are as follows:

1) the syllable pitch specified by a rather narrow range and l e v e l changes in the fundamental pitch, and also a falling, rising, or rising-falling direc-

tion of tone.

In the Latvian standard language, in which the falling, drawling and broken syllable toneme is contrasted, the falling toneme is more distinctly realized (with a falling direction of the fundamental pitch). Be it otherwise, it would coincide both, with a broken and drawling syllable toneme. In the system of two types of syllable toneme, functioning in the Ziemeri sub-dialect, it is essential that the falling syllable toneme should not coincide with the broken syllable toneme. Therefore, the falling syllable toneme is subject to greater variations - obtaining the characteristics similar to the drawling syllable toneme of the standard language;

2) a very distinct syllable toneme specified by a wide range and a c u t e changes in the fundamental pitch, and, likewise, a snift falling tone.

There are cases when the direction of the tone is sniftly falling in the first half of the vowel, but rising - in the second half of it without reaching the maximum frequency.

The second type of the syllable toneme is specified by a decrease in the regularity of vocal-chords vibrations, the so-called glottalization, pointed also out by A.Ābele, A.Laua, I.Lehiste, M.Neilande, M.Vecozola, and others. The design of the irregular vibrations of this kind bears resemblance to the broken toneme (the so-called stod) of the Danish.

Until all the variants of the High Latvian dialect are not examined experimentally, it is disputable in which part of the vowel with a broken syllable toneme, in Latvian sub-dialects, the loss of regular vibrations occurs, or where there is a complete discontinuation (break) of voice.

The analysis of spectrum dynamics of the vowels in the Ziemeri sub-dialect proves the weakening, or even fading of formants, in the case of the broken syllable toneme, e.i. an a c u t e change in the dynamic design can be observed. In the sub-dialect under consideration, acute changes of the spectrum design are observed in the transition part of a diphthong. Sometimes the fading of formants is somewhat delayed, e.i. it occurs at the beginning of the second element. After the break some spectrograms all the three constituent parts of the formant are distinctive enough. In the monophthongs with a broken syllable toneme, of the Ziemeri sub-dialect, irregular vibrations or a complete disappearance (break) of voice occurs at about the end of a third part, or in the middle of the syllabic element.

When investigating syllable toneme by auditive methods, Latvian philologist A. Breidaks expressed a view that in many variants of the High Latvian dialect, the disappearance, or acute shanges in voice occur in the first (prolonged) element of a diphthong or a diphthongal combination having a broken syllable toneme /1/. A. Breidaks refers also to the research made by A.Ābele and M.Lepika /2/, who had analysed by auditive method the texts of the Jaunlaicene, Jaunroze and Veclaicene sub-dialects, which are adjacent to that of Ziemeri. Yet in another contribution /3/ concerned with the Alūksne sub-dialect, which is also adjacent to its Ziemeri counterpart, on the basis of the experimental data obtained by kymographic analysis, A.Ābele states that acute changes occur in the final part of a diphthong. In the course of the experimental investigation of the other three variants of the High Latvian dialect, I.Martinsone observed acute changes in the middle or the second element of a diphthong uttered with a broken syllable toneme /4/. Summing up the statements mentioned above, we may conclude that the experimental research of the vocalism used in different variants of the High Latvian dialect proves the occurence of acute changes, or disappearance of voice, in the case of the broken syllable toneme, in the transition part,or the beginning of the second element of a diphthong. The fact is contradicted by auditory percèption and, therefore, must be subjected to a careful experimental test. Experimental research of uninvestigated variants of the High Latvian dialect is left for the future.

In the spectra of the vowels with a falling syllable toneme, a distinct relevant feature - acute changes - are absent. The changes in the dynamic design of the spectra are l e v e l.

These occurences testify to the fact that, in the Ziemeri sub-dialect,the broken syllable toneme is a marked element of opposition with regard to the unmarked toneme of the first type. This kind of toneme corresponds to the conventional term used for the broken syllable toneme. The first type of toneme is conventionally called falling, it would be more precise, from a phonological view point, to call the first type - unbroken or level.

According to dynamics and intensity direction in the vowels of the Ziemeri sub-dialect, two types of syllable pitch can be contrasted:

1) the syllable toneme specified by l e v e l changes in intensity and also by a rising-falling direction of intensity;

2) the syllable toneme specified by a c u t e changes in intensity, and, also, by an acute falling or rising-falling direction of intensity. In some cases intensity may have a quick fall in the first half and a rise in the second half of a vowel without ascending to a maximum intensity of the first half of a vowel.

We may conclude that in both types of syllable toneme a rising-falling intensity occurs, consequently, intensity direction (as well as the direction of the fundamental pitch) is of no significance in the differentiation of the types of syllable toneme in the Ziemeri sub-dialect. Both types are contrasted to each other by the presence or absence of a specific prosodic distinctive feature - an a c u t e or l e v e l characteristics of intensity changes (as well as the fundamental pitch changes).

Depending on the syllable toneme in the Ziemeri sub-dialect, long monophthongs differ as to their duration: long monophthongs with a falling syllable toneme (M_f) exceed long monophthongs with a broken syllable toneme (M_b) in their duration. An avrage correlation is: $M_f : M_b : M = 1{,}7 : 1{,}2 : 1$ (M - short monophthongs).

In the sub-dialect the duration of diphthongs is close to that of long monophthongs.

The differentiation of the syllable toneme types in the Ziemeri sub-dialect is based on the spectrum, fundamental pitch, intensity and duration of vowels. Each of these parameters plays a certain role in differentiating toneme. For example, the acute changes in the fundamental pitch and intensity, the decline in the tembre of monophthongs (reflected by the lowering of formants in a spectrum) and the reduced duration may signal the presence of a broken syllable toneme. Yet not a single parameter functions as the only, basic and reliable indicator. The spectrum, fundamental pitch, intensity and duration seem to compensate each other. It is credible that in certain phrases or intonation patterns the decisive role is played by one or the other of these distinctive features (for instance, it may be considered by preliminary observations that in interrogative phrases pitch, to a certain extent is deprived of its ability to differentiate syllable tonemes).

The syllable tonemes of monophthongal or diphthongal syllables, in fact, do not bear distinction among them - their distinctive features fully coincide. Judging by auditory perception, the distinctive features of the same kind are present in diphthongal clusters,which were not investigated by us.

We may conclude that both types of syllable toneme are contrasted to each other by the presence or absence of the specific prosodial feature - a c u t e or l e v e l changes in the fundamental pitch, intensity and spectrum. See also some illustrations (Fig. 1, 2, 3, 4) of the fundamental pitch and intensity of vowels with the both types of syllable toneme.

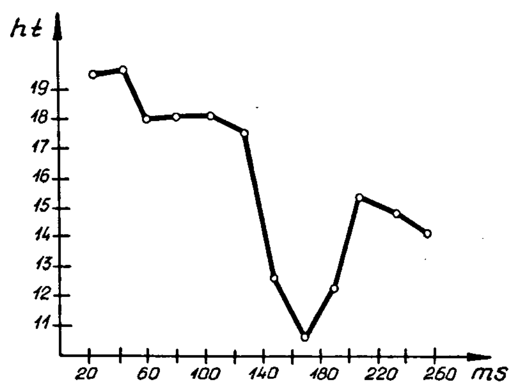


Fig.1. Diphthong iê in the word spiêra

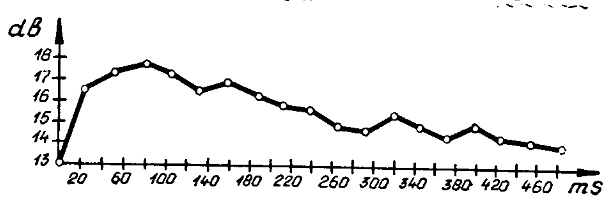


Fig.2. Monophthong ì in the word pìra

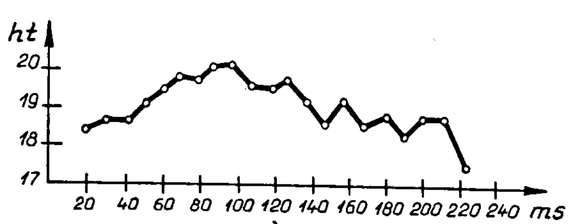


Fig.3. Diphthong ùo in the word kùosu

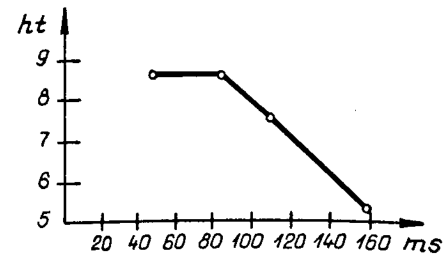


Fig.4. Monophthong â in the word âstu

We consider that, from the phonological view point, it would be more apt to call the two types of syllable toneme - a c u t e (or broken) and l e v e l (e.i. unbroken).

REFERENCES

1. A.Breidaks. Latgalisko izlokšņu prosodijas jautājumi. // Grām.: Veltījums akadēmiķim Jānim Endzelīnam 1873.-1973., Rīga: Zinātne, 1972, 89.-108.lpp.
2. A.Ābele un M.Lepika. Par Apukalna izloksnēm. // Filologu biedrības raksti, 1928, VIII, 19.-49.lpp.
3. A.Ābele. Alūksnes izloksnes intonācijas. // Filologu biedrības raksti, 1930, X, 80-91.lpp.
4. J.Martinsone. Vārkavas, Pildas un Zvirgzdines pagasta intonācijas. // Filologu biedrības raksti, 1934, XIV, 143.-165.lpp.

MODELLING SWEDISH SEGMENT DURATION

ROLF CARLSON AND BJÖRN GRANSTRÖM

Department of Speech Communication and Music Acoustics,
Royal Institute of Technology, Box 70014,
S-100 44 Stockholm, Sweden.

ABSTRACT

The durational properties of consonants have been studied for Swedish in the context of a speech data base of read sentences. We have developed a system to access a speech data base in an effective manner by means of rules. These rules can also be used to describe models that can be tested against the data. Some durational effects such as inherent duration and stress and quantity effects have been verified. Durational attributes of boundaries play an important role in a complete account of prosody. Syllable, morph, word and phrase boundaries have to be taken into account. The needs for larger speech data bases are obvious when finer details are going to be studied and described. Our main objective in this paper has been to illustrate the method and to show the power of the approach.

INTRODUCTION

Durational data has been reported for several languages and also formulated into coherent rule systems. Only Swedish data and models will be discussed and referred to in this paper. An expanded version of this paper also includes data for American English /1/.

A speech data base of read Swedish sentences has been created and methods to search this data base by means of rules are also reported. The prosodic analysis of Swedish in this paper consists of both duration analysis of consonants and testing of duration models. The models are based on a general structure proposed by Klatt /2/.

THE SWEDISH SENTENCE DATA BASE

The speech data base in our example consists of 150 Swedish sentences, containing about 5000 phonemes, read by one male speaker. The first step in creating the data base was to record and label speech. In our system, speech data is stored in sentence-sized files. Our text-to-speech system is used to phonetically transcribe the utterances /3/. This transcription is edited to match the pronunciation as well as possible (Figure 1). It is a matter of discussion how detailed this transcription should be. We are aiming at a relatively broad phonemic transcription. We believe that the broader transcription makes it easier to use the data base to discover and study phonetic variations of certain kinds. An example is devoicing of voiceless sounds in voiced contexts which appears to be a graded phenomenon rather than an allophonic selection. Stress and word-tone is marked by special signs. Additional markers indicating e.g. syntactic boundaries and emphasis can be added to the transcription if needed.

The phonetic transcription is used by an automatic segmentation program, /4/, to distribute the phonetic labels along the wave form. The segmentation program gives an estimate of the time position of each phoneme. Segmentation of speech in phone-sized parts in an unambiguous way is a classical problem, possibly without a solution. When a number of persons are contributing to the data base, it is important that the same criteria are used throughout. An attractive alternative is to leave the segmentation to a self-consistent algorithm. The accuracy of the present program is, however, not sufficient.

When a detailed analysis should be done, the labels have to be checked and corrected. This is done by means of a wave form editor program, which is a general purpose program for labelling and editing sampled files. By means of the joystick,

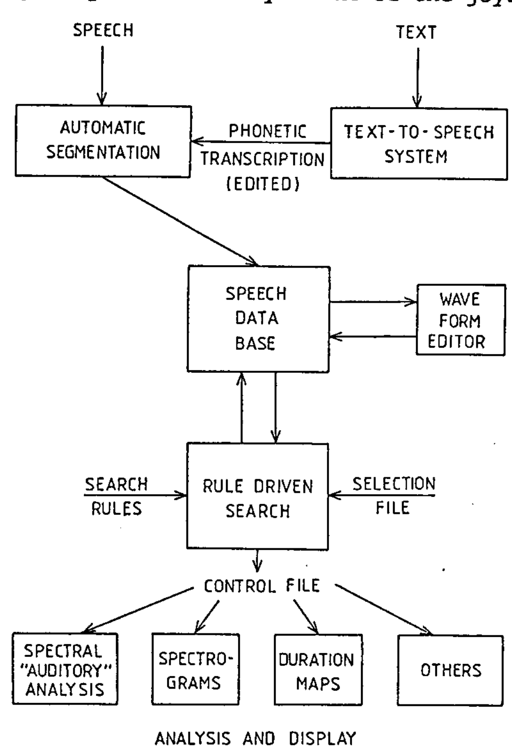


Figure 1. Block diagram of the rule controlled data base environment.

samples can be labeled or labels can be changed. The labels are stored in label files which are used by all following programs. During the editing, the program can suggest good positions for labels. This is done by an automatic procedure that places the cursor at zero crossings or at the closing time of the glottal source. These features make the program fast, interactive, and user-friendly.

Figure 2 shows a spectrogram of a sentence pronounced by the same speaker that we used in the KTH data base. The label names and positions can be seen at the top.

Labelling speech is often a difficult task. In many cases no obvious segment boundary can be found. This is especially the case in sequences of segments sharing the same manner of articulation. In many of these cases the labels have to be set according to some conventions that can be coupled to acoustic events. Even though the label position can sometimes be regarded as ambiguous or even meaningless it is important to always supply it. By having a labeled data base we have the possibility of identifying sounds in a specific context for further analysis which is not crucially dependent on the exact label position.

RULE-DRIVEN SEARCH

The data base is accessed by means of rules. By a brief rule statement, speech segments meeting the specified contextual conditions can be identified. The rule structure is similar to the notation used in generative phonology and is also used in our text-to-speech project.

The rules operate on the transcription and are used to insert a '*' symbol in front of the phoneme to be analyzed and to give it a set of parameter values. These parameters can be used to specify the time position for each phoneme, the duration of the phoneme, the stress level, or any information that can be derived from the phonetic transcription or the durational information in the label file. Table I gives an example of a simple rule system to find all vowels and to classify them depending on stress level and phonological length. If the vowel precedes an unvoiced stop it is given a higher classification number. The result of the analysis is shown in Figure 3.

Table I. Rule system to find and classify vowels.

```
insert * in front of vowels
01.00: ^ *·/ & <VOWEL>
save vowel durations in the *; give all vowels class 1
02.00: * ^ <DUR:=Y,CLASS:=1> / & <VOWEL,Y:=DUR>
give class 2 to short vowels with primary stress
04.00: * ^ <CLASS=2> / & <VOWEL,STRESS,1STRESS,-TENSE>
give class 3 to long stressed vowels
05.00: * ^ <CLASS=3> / & <VOWEL,STRESS,TENSE>
add 3 to the class if vowels are before voiceless stops
07.00: * ^ <CLASS=CLASS+3> / & <VOWEL> <STOP,-VOICE>
```

It is a well known fact that a vowel is shortened when followed by an unvoiced stop. However, we find support for a strong shortening effect only in short stressed vowels while the other two categories have a minor shift in duration. Also we find that the unvoiced stops have a much higher long/short ratio than other consonants in our data.

A special feature of the system is that the rule notation itself is a powerful tool to describe a model such as a text-to-speech system. The model prediction can, thus, be immediately compared with the actual data during the data base search. Some examples will be given in this paper.

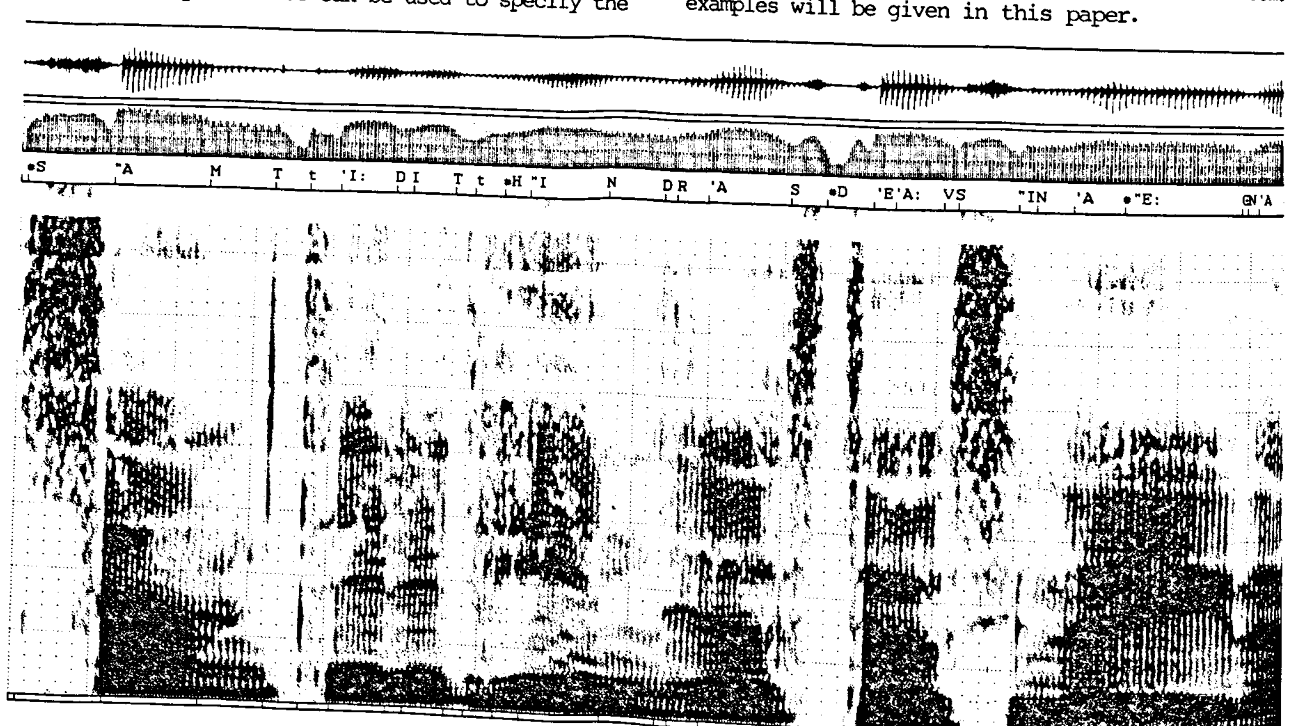


Figure 2. Spectrogram of the sentence "Samtidigt hindras de av sina..."

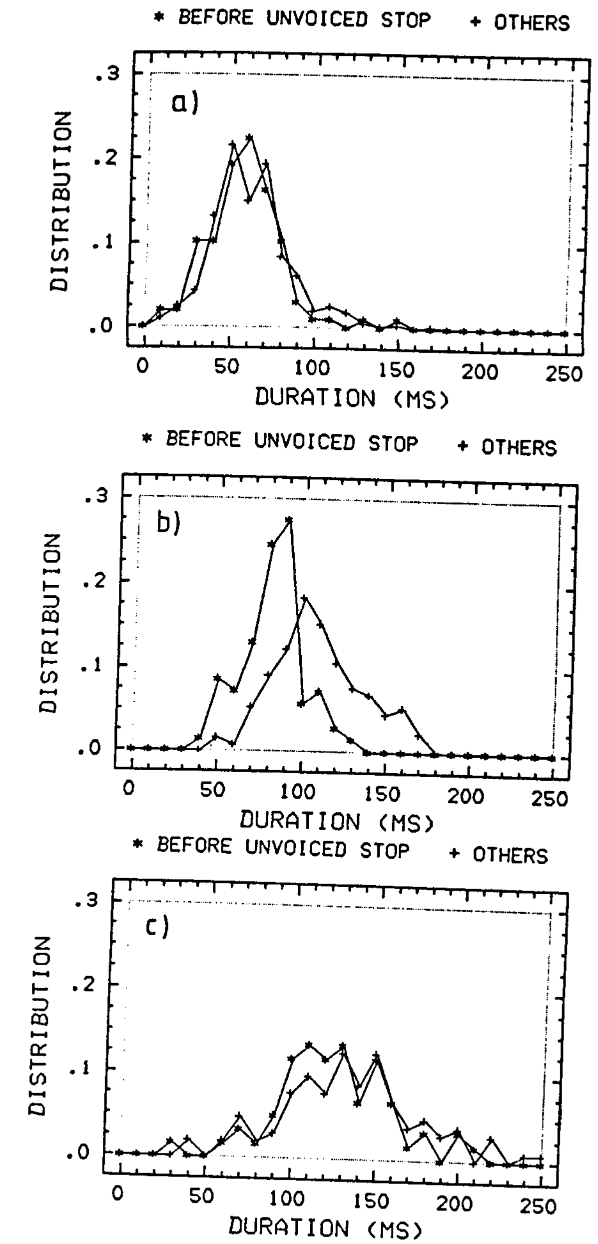


Figure 3. Influence of stop consonants on the preceding vowel. (a) unstressed vowels, (b) stressed short vowels, (c) stressed long vowels.

DURATION ANALYSIS OF THE KTH DATA BASE

Durational variation in consonants depend on several factors including consonant type, stress and immediate phonetic context. As a reference point the mean and SD for all 2917 consonants was found to be 60 ms. and 34 ms. respectively. All clause-initial and clause-final consonants are excluded from the analysis. Some of the variation can be taken care of by splitting the material in different groups. The decrease in SD is used as a measure of the predictive power of the categories.

At first the consonants are divided into three major classes: unstressed, stressed and stressed long consonants. In the present analysis a consonant is defined to be stressed if it is followed by a primary stressed vowel. A consonant is regarded as stressed long if it immediately follows a primary stressed short vowel. These definitions need to be modified as will be seen in the following analysis.

In Table II the number of occurrences, mean and SD for the subcategories are presented. To take into account each consonant's typical length, we calculated the mean for each consonant in the three categories and estimated the variation (SD*) in relation to these means. The result is interesting in the context of a text-to-speech system. If we give each consonant three typical duration values we will get a prediction that only takes care of 25 percent of the original SD.

Table II. Mean and SD for different Swedish consonant classes.

	N	mean (ms.)	SD (ms.)	SD* (ms.)
unstressed	1717	54	29	25
stressed short	806	62	26	21
stressed long	394	83	33	30
all consonants	2917	60	34	25

The next step in our analysis is to break down our data into more specific subgroups. We have divided the material into word-initial and word-medial or -final consonants (Table III).

Table III. Mean and SD for Swedish consonants.

	INITIAL			MEDIAL or FINAL		
	MEAN	SD	N	MEAN	SD	N
UNSTRESSED						
$\underline{C}$	53	19	177	53	26	927
$\overline{C}C$				54	22	169
$\overline{C}\underline{C}$				53	19	216
STRESSED						
$\underline{C}$'V	69	21	406			
$\overline{C}C$'V	63	21	171			
$\overline{C}\underline{C}$'V	49	16	188			
'V$\overline{C}$:				92	32	181
'V$\overline{C}$:C				75	24	213
'V$\overline{C}$:$\underline{C}$				61	21	206

The long consonants have the expected increased duration and this increase is maintained even if the consonant is followed by another consonant. Even the second consonant following the long consonant is longer than the unstressed consonant. Therefore, to be able to do a correct prediction of duration in Swedish, we have to know the syllabic structure which is difficult to derive even from a theoretical point of view.

ANALYSIS IN THE CONTEXT OF A MODEL

We have so far discussed some broad analysis of the consonant duration in the present KTH data base. As mentioned earlier the data base is too small for very specific analyses. Even inherent duration, according to the definition above, is hard to measure reliably. Swedish words often end with consonants and to make a natural data base with a statistically reasonable frequency of single word-initial stressed consonants preceded by vowels demands a considerably larger corpus.

We have chosen to approach the material from a different point of view. We have implemented the rule system presented by Klatt (1979) as part of the data base search. This makes it possible to test the predicted duration against the measured. The rules are based on the concepts of inherent duration, minimal duration and a correction factor. Only a few of the rules are applicable for our purpose. The rule numbers refer to the rule system in Klatt´s work.

Find inherent duration INHDUR and minimal duration MINDUR in a phoneme-specific table. Set adjustment parameter : PRCNT=1.0

Rule 6. Noninitial-consonant shortening. Consonants in nonword-initial position are shortened by : PRCNT = PRCNT *. 85

Rule 7. Unstressed shortening. Unstressed segments are half again more compressible than stressed segments. Then both unstressed and 2-stressed segments are shortened: MINDUR=MINDUR/2 and PRCNT=PRCNT*.7

Rule 10. Shortening in clusters. Segments are shortened in consonant-consonant sequences (disregarding word boundaries, but not across phrase boundaries).
consonant preceded by consonant: PRCNT = PRCNT * .70
consonant followed by consonant: PRCNT = PRCNT * .70

Rule xx. Long consonants after primary stress were adjusted according to the rule: PRCNT = PRCNT * 2

Calculate the resulting duration:
DUR = (INHDUR-MINDUR)*PRCNT + MINDUR

As a starting point the rules were implemented and the inherent duration and the minimal duration were estimated from the predictions and actual data. In a sequence of test runs these values were optimized. The results are presented in Table IV.

Table IV. Inherent and minimal duration for Swedish

	INHDUR	MINDUR		INHDUR	MINDUR
b occl	65	50	f	90	60
d occl	55	40	s	100	50
rd occl	55	40	rs	100	50
g occl	50	40	sh	95	60
p occl	65	50	h	90	20
t occl	50	40	v	50	40
rt occl	50	40	j	65	35
k occl	50	40	r	50	30
m	65	50	l	65	40
n	70	40	rl	65	40
rn	70	40			
ng	80	50			

The first test showed a SD of 23 ms., which should be compared to the initial 34 ms. without consonant-specific adjustments and 25 ms. with the three category classification. The improvement is minor and not statistically significant. It is however unfair to claim that the rule system has little or no positive features. What is missing is to adjust the rules to the syllabic nature of the Swedish language and to include the important phrase rules. If a simple stripping of unstressed endings and prediction of secondary stress in compounds together with a few other rules were added the SD decreased to 20 ms. The comparison of the measured and predicted consonant durations can be visualized in graphical form. We still get gross errors at phrase boundaries. Excluding these we find the quite acceptable SD of 13 ms.

CONCLUSION

We have developed a system to access a speech data base in an effective manner by means of rules. These rules can also be used to describe models that can be tested against the data. This method has been used to study the durational structure of Swedish. Some durational effects such as inherent duration and stress and quantity effects have been verified. Durational attributes of boundaries play an important role in a complete account of prosody. Syllable, morph, word and phrase boundaries have to be taken into account. The need for larger speech data bases is obvious when finer details are going to be studied and described. Our main objective in this paper has been to illustrate the method and to show the power of the approach. The current system enables us to test hypotheses and to transform the gained knowledge to our text-to-speech system or speech recognition system in a fast and effective manner.

ACKNOWLEDGEMENTS

Part of the work was supported by The Swedish Board for Technical Development (STU) Contract No. 84-3667.

REFERENCES

/1/ Carlson, R. and Granström, B.: A search for durational rules in a real-speech data base, Phonetica, Vol. 43:140-154 (1986).

/2/ Klatt, D. K.: Synthesis by rule of segmental durations in English sentences, in Frontiers in Speech Communication Research, ed. B. Lindblom and S Öhman (Academic, New York 1979).

/3/ Carlson, R., Granström, B., and Hunnicutt, S.: A multi-language text-to-speech module, Conference Record, IEEE-ICASSP, Paris (1982).

/4/ Blomberg, M. and Elenius, K.: Automatic time alignment of speech with a phonetic transcription, STL-QPSR 1/1985:37-45 (1985).

STATISTISCHE ZEITPARAMETER DER GESPROCHENEN SPRACHE

GOTTFRIED MEINHOLD

Sektion Sprachwissenschaft
Friedrich-Schiller-Universität
DDR 6900 Jena

ABSTRACT

Eine empirische Untersuchung an etwa 100 Sprechern gibt Aufschluß über die Aussagekraft sog. makrozeitlicher Parameter, mit deren Hilfe wichtige phonostilistische Besonderheiten der gesprochenen Sprache beschrieben werden können. Die Untersuchung reicht von expressionsarmen Nachrichten des Rundfunks bis zu sprechkünstlerisch gestalteten Texten (Lyrik, Ballade, Prosa, Sprechtheater). Einige korrelative Beziehungen zwischen den Parametern werden festgestellt.

Neben der mikrozeitlichen Struktur gesprochener Texte, wie sie sich aus Lautdauermessungen ergibt oder in Silbendauer oder Redetaktdauer zum Ausdruck kommt, ist eine makrozeitliche Struktur aufschlußreich, deren wichtigste Parameter geeignet sind, die Verschiedenartigkeit der gesprochenen Form von Texten zu charakterisieren sowie die temporalen Verlaufsformen sog. freier lautsprachlicher Äußerungen zu kennzeichnen.
Solche Parameter sind (1) die mittlere Sprechgeschwindigkeit V_o einer Sprachäußerung, ausgedrückt in Silben/s, (2) die Streuung der Syntagmageschwindigkeiten, die aus der Verteilung der mittleren Geschwindigkeit jener Textabschnitte zu ermitteln ist, welche nicht durch eine Pause getrennt werden (Pausensyntagmen), (3) die mittlere absolute Geschwindigkeitsfluktuation (Gf), berechnet als mittlerer Betrag der Geschwindigkeitsdifferenzen zwischen zwei aufeinanderfolgenden Pausensyntagmen, (4) der Pausenzeitquotient, berechnet aus der Gesamtzeit des Textes (t) und der reinen Sprechzeit (t_o), also $PZQ = t/t_o$, (5) die mittlere Pausendauer ($\bar{t}_p$) des Textes, (6) die auf 1oo Silben bezogene Pausenhäufigkeit (P_n). - Prosodische Parameter, die mit den angeführten Zeitparametern in engem Zusammenhang stehen, sind die mittlere Akzentdichte, ausgedrückt durch den mittleren Akzentabstand sowie die rhythmische Kontextentropie, die aus den Verbundhäufigkeiten der Redetaktkette des Textes berechnet wird.
Eine größere Anzahl Textproduktionen und -reproduktionen wurden auf die genannten Parameter hin analysiert; unter den Sprechern befanden sich auch solche, die Deutsch als Fremdsprache studieren (Muttersprache Russisch, Ukrainisch).
Die Sprechgeschwindigkeit V_o eines Textes weist in verschiedenen Sprachäußerungen die folgenden Mittelwerte auf (Silben/s):

	Mittelwert	Streuungsbereich
Nachrichten	5,7	5,4-6,1
Belletristik	5,4	4,6-6,1
Sprechtheater	4,97	3,5-6,1
Lyrik/Ballade	3,7	3,1-4,4
freie Äußerung	5,9	5,5-6,1

Die freie Redeäußerung betraf die Verbalisierung einer Bildgeschichte. - Statistisch signifikant ist lediglich der Unterschied zwischen 'Lyrik/Ballade' und allen anderen Gruppen; es gibt also hinsichtlich V_o nur eine Differenzierung zwischen sog. versgebundener und nichtversgebundener Rede. Ermittlungen an Sprechern russischer und ukrainischer Muttersprache ergaben mit 5,8 Silb./s für die freie Äußerung dem Deutschen angenäherte Werte.

PZQ (s.o.) drückt den (relativen) Anteil der Pausenzeit (t_p) aus. PZQ hat die Tendenz, bei abnehmender Sprechgeschwindigkeit zu wachsen.

	Mittelwert (PZQ)	Streuungsbereich
Nachrichten	1,21	1,13-1,26
Belletristik	1,40	1,22-1,81
Sprechtheater	1,38	1,23-1,59
Lyrik/Ballade	1,42	1,23-1,79
freie Äußerung	1,62	1,3o-1,75

Bemerkenswert ist der kleine Streuungsbereich der Nachrichten; mit der Zunahme der Mittelwerte wächst der Streuungsbereich nach oben. Es hat sich gezeigt, daß bei den reproduzierten Texten PZQ wächst, wenn die Expressivität des Sprechausdrucks zunimmt. In der freien Äußerung spiegelt sich dagegen im hohen PZQ vor allem der Planungszeitaufwand der Sprechproduktion wider. Phonostilistisch deutet sich die prosodische Normierung und geringe Variabilität der Nachrichten an.

Die Streuung der Pausensyntagmageschwindigkeiten (H) und die Geschwindigkeitsfluktuation wurde nur an einem Teil des Corpus ermittelt (Prosa, Sprechtheater und freie Äußerungen):

	H (in bit)	Gf (Silb./s)
Prosa	2,7o74	o,93
Sprechtheater	3,o131	1,32
freie Äußerungen	3,1	1,35

Beim Vergleich mit anderen Zeitparametern erweist sich, daß sich die Gruppen 'Prosa' und 'Sprechtheater' nur hinsichtlich H und Gf signifikant voneinander unterscheiden lassen. Bemerkenswert ist die geringe Differenz zwischen 'Sprechtheater' und 'freien Äußerungen'.

Von den Pausenparametern ist die mittlere Pausenzeit ($\bar{t}_p$) lediglich geeignet, weit auseinanderliegende Textklassen (Lyrik/Ballade einerseits mit $\bar{t}_p$ = o,8 s, Lesungen von Prosa mit $\bar{t}_p$ = o,6 s) zu differenzieren. Die bezogene Pausenhäufigkeit P_n signalisiert demgegenüber rhythmische Strukturverschiedenheiten, die idiolektisch oder situativ variabel sind. Neben P_n kann auch der mittlere Pausenabstand P_a in Silben berechnet werden.

	P_n	P_a
Nachrichten	5,8	17,3
Belletristik	11,5	8,3
Sprechtheater	9,4	1o,6
Lyrik/Ballade	15,o	6,7
freie Äußerung	11,2	8,2

Nachrichten und Lyrik/Ballade stellen die extremen Gruppen dar. Betrachtet man die Pause als syntagmatisches Gliederungsmoment, das auch Recodierschritte beim Verstehensprozeß markiert, so geben die Werte des Sprechtheaters einen Hinweis auf die optimale Recodierschrittlänge.

Weitere Differenzierungsmöglichkeiten zeigen sich, wenn man Pausen zwischen den Sätzen (SZP) und Pausen innerhalb der Sätze (SP) gesondert betrachtet. Sprechtheater (DR) und Prosalesung (PR) unterscheiden sich in dieser Hinsicht signifikant voneinander.

	Satzzwischenpausen		Satzpausen	
	$\bar{t}_{szp}$	P_n	$\bar{t}_{sp}$	P_n
PR	1,5	2,5	o,53	6,4
DR	1,1	3,5	o,69	5,3

Ein Quotient aus den Werten für die Pausendauer (PQ = $\bar{t}_{szp}/\bar{t}_{sp}$) spiegelt das zugunsten der SZP verschobene Verhältnis in der Prosa wider (PR: 2,87; DR: 1,69). - Die Abhängigkeit zwischen P_n und V_o ist geeignet, die phonostilistisch unterscheidbaren Textklassen 'Lyrik/Ballade', 'Nachrichten' und 'Belletristik' zu unterscheiden; die Abhängigkeit $\bar{t}_p$ von V_o leistet dies nicht. - Die Erwartung, daß bei zunehmender Sprechgeschwindigkeit P_n und $\bar{t}_p$ sich verringern, gilt allenfalls für einen bestimmten Bereich emotionaler Neutralität. In einer Leseprobe des Corpus wird als Ausdruck für resignative Haltung hohe Sprechgeschwindigkeit in Verbindung mit großer mittlerer Pausendauer und Pausenhäufigkeit realisiert (V_o = 6,13 Silb./s; $\bar{t}_p$ = 1,18 s; P_n = 11,o Paus/1oo Silb.) Es kommen also Tendenzen zu positiver wie negativer Korrelation vor. Deutlicher zeigt sich die Tendenz zu einer negativen Korrelation zwischen V_o und P_n in den freien Äußerungen, dagegen scheint hier zwischen V_o und $\bar{t}_p$ keinerlei korrelative Beziehung zu bestehen. Erwartungsgemäß liegt aber eine korrelative Beziehung zwischen PZQ und $\bar{t}_p$ vor.

Nichttemporale prosodische Paramter, die mit den erwähnten Zeitparametern in engem Zusammenhang stehen, sind Akzentstruktur bzw. rhythmische Struktur eines Textes, ausgedrückt durch den mittleren Akzentabstand als Maß für die 'Akzentdichte', sowie die rhythmische Kontextentropie, die aus den relativen Verbundhäufigkeiten der Redetaktklassen berechnet wird. Mit wachsender Akzentdichte und zunehmend kürzeren Redetakten erhöht sich der rhythmische Ordnungsgrad (verringert sich die rhythmische Entropie) und nehmen die Sprechgeschwindigkeit sowie die Pausenhäufigkeit ab. All dies sind Kennzeichen eines wachsenden expressiven Spannungsgrades.

Die temporale Makrostruktur gesprochener Texte ist geeignet, einen Beitrag für eine genauere Beschreibung phonostilistischer Gegebenheiten, jedoch auch individualtypischer Erscheinungen der Sprechweise zu leisten. Außerdem liefern ihre Zeitparameter objektive Angaben für die 'fluency of speech' und somit für die Bestimmung des qualitativen Niveaus der (nicht nur phonetischen) Sprachbeherrschung. Darüber hinaus sind sie für psycholinguistische bzw. sprachpsychologische und sogar für kriminologische Ermittlungen aufschlußreich.

LITERATUR

Gajdučik, S.: Zur phonostilistischen Differenzierung der gesprochenen Hochsprache. In: Zeitschr.f.Phonetik, Sprachwiss.u.Kommunikationsforschung, 25 (1972) 47-57

Meinhold, G.: Allgemeine phonetische Probleme der Sprechgeschwindigkeit. In: Zeitschr.f.Phonetik, Sprachwiss.u. Kommunikationsforschung 25 (1972) 491-5o5

Raabe, M.: Zur Charakterisierung sprecherischer Verlaufseigenschaften durch phonostilistische Merkmalskomplexe. In: Wiss. Zeitschr. d. Friedrich-Schiller-Universität Jena, Gesellschaftswiss.R. 34 (1985) 97-1o6

SOME OBSERVATIONS ON THE TIMING OF FO-EVENTS

Bertil Lyberg

Swedish Telecommunications Administration
Technology Department, S-123 86 Farsta, Sweden.

ABSTRACT

The present study examines the effects which final consonants have upon the timing of the fundamental frequency contour in words carrying the sentence accent in Swedish. Monosyllabic test words containing both phonologically long and short vowel segments are placed in initial and final utterance positions. Results show that the timing of the FO-events that signal the sentence accent is dependent on whether the consonant following the vowel is voiced or not, especially when the vowel is phonologically short. The fundamental frequency fall in the case of a short vowel followed by an unvoiced consonant has to occur earlier than in the other cases in order to get the frequency fall within the vowel segment, otherwise the prosodic information will get lost.

INTRODUCTION

The fundamental frequency contour of an utterance is heavily influenced by the segmental composition. In investigations about the fundamental frequency contour utterances built up of only sonorants are often used [1] and thereby the influence of constrictions in the vocal tract is avoided or at least diminished. It is assumed that the fundamental frequency contour obtained in such a way will reflect some basic pattern that is perturbed by the segmental composition in ordinary utterances. Most studies of the fundamental frequency contour are dealing with overall patterns i.e. in which syllables maxima and minima will occur in the segmental flow [1] [2]. In models for generating an accurate fundamental frequency contour it is also necessary to take into account how the location of the maxima and minima is affected by the segmental composition. The exploration of such effects in greater detail is hopefully of great importance for the generation of synthetic speech with a naturalness and intelligibility that is acceptable in different types of communication systems. A systematic mapping of the variations of the locations of the extremes owing to the segmental setup can also be expected to provide some insight into certain aspects of the mental processes underlying the temporal organization of spoken language.

In the present investigation the fundamental frequency contour associated with sentence accent is studied in greater detail. The syllable structure and word position are systematically varied and the effects on the location of the extremes are examined.

Some fundamentals of Swedish prosody

The fundamental frequency contour of a monosyllabic word carrying sentence accent will in phrase final position have a maximum point in the vowel followed by a minimum point in the vowel or the following consonants. For a more detailed description of the fundamental frequency contour in different positions of Swedish utterances, see e.g. Bruce [1] and Lyberg [3]. The fundamental frequency manifestation associated with the signalling of sentence accent in Swedish seems to be very similar to the corresponding frequency manifestation in American English according to e.g. Pierrehumbert [2]. There are, however, discrepancies in the interpretation of the underlying parameters, and in the terminology used by the two authors Bruce and Pierrehumbert.

Two degrees of quantity are distinctive in Swedish, the short/long distinction. There is also a complementary distribution of phonological length between vowels and consonants in stressed syllables. A long vowel is in stressed syllables followed by a short consonant and a short vowel by a long consonant [4].

EXPERIMENTAL DESIGN

Speech material

A set of utterances containing one and three lexical main stresses was constructed. The test word was in the case of three-word sentences placed in both initial and final positions and the sentences were pronounced either with the test word in focus or with a "neutral" stress pattern i.e. with a conscious effort of the speaker to avoid junctures and contrastive stresses.

The test word was build up of both phonologically short and long vowel segments in order to elucidate the interaction between the signalling of the quantity distinction and the fundamental frequency contour. In addition to that the surrounding consonants were varied in a systematic way so that both voiced and unvoiced consonants occurred in postvocalic position. The test words were always monosyllables and may be considered as nonsense words. The seminonsense three-word utterances where built up of both nonsense (test words) and semantically non-anomalous words (always /såg/ saw in English).

The inventory and syntactic structure of the test sentences are presented in Table I. The sentences were read in ten randomly ordered sequences by a trained phonetician.

TABLE I

Sentences	Syntactic structure
{Dad, Dadd, Dat, Datt}	S – NP – N
{Dad, Dadd, Dat, Datt} säg {dad, dadd, dat, datt}	S → NP (N), VP (V, NP (N))

+ A phonologically short vowel is in the orthography followed by two consonants.

Measurements

The duration of the vowel segment in the test words was measured. The duration of the vowel segment is defined as the interval between the release of the consonant preceding the vowel (always /d/; a rapid increase of intensity) and the occlusion of the following consonant (always /d/ or /t/; a rapid decrease of intensity).

The fundamental frequency was measured in nine equally spaced points of the vowel segment in the test words, at the beginning, at the end, and at seven points within the vowel segment.

OBSERVATIONS

The fundamental frequency contour is in figs. 1 and 2 shown for the final position of the three-word utterances. The diagram in fig.1 shows the fundamental frequency contour when the utterance is pronounced with a "neutral" intonation pattern and the diagram in fig.2 the frequency contour when focus is assigned to the final position of the utterance. Every point in the diagrams represents a mean value of ten recordings ot the same utterance. The maximum point and the following minimum point of the fundamental frequency contour will occur within the vowel segment no matter whether the following consonant is voiced or not and whether the vowel in question is phonologically long or not. The fundamental frequency contour after the minimum point is, in the case of a phonologically long vowel followed by an unvoiced consonant or a short vowel followed by a voiced consonant, more or less truncated in comparison to the frequency curve in the case of a long vowel followed by a voiced consonant. When a short vowel is followed by an unvoiced consonant the fundamental frequency fall will occur about 20 to 30 msec. earlier in the vowel segment in comparison to the other cases.

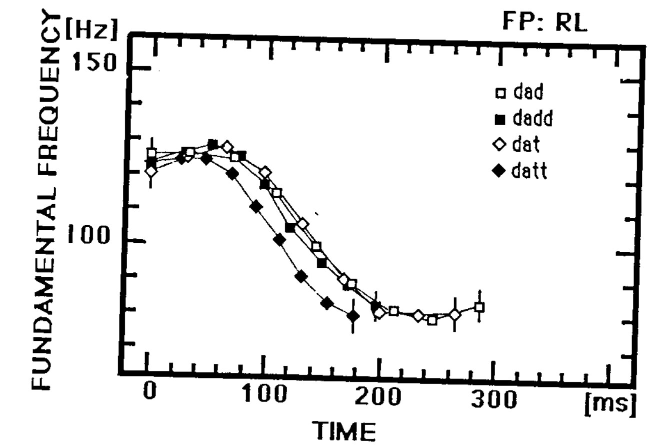


Fig. 1 The fundamental frequency contour of the vowel for different test words in final position of three-word utterances. The utterance is pronounced with a "neutral" intonation pattern.

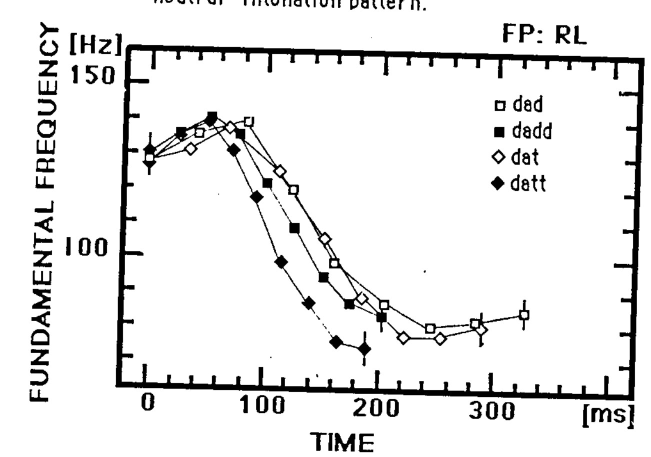


Fig. 2 The fundamental frequency contour of the vowel for different test words in final position of three-word utterances. Focus is assigned to the final position.

In the diagram in fig.3 the fundamental frequency contour is shown for the one-word utterances. The diagram shows that mainly the same timing difference of the fundamental frequency fall is apparent in these utterances. When a short vowel is followed by an unvoiced consonant the frequency fall will happen earlier in comparison to the other cases. A comparison between the fundamental frequency fall of a final test word carrying sentence accent in a three word utterance and a comparative test word in a one-word utterance shows that the fundamental frequency fall in the one-word utterance will happen later than in the three-word utterance.

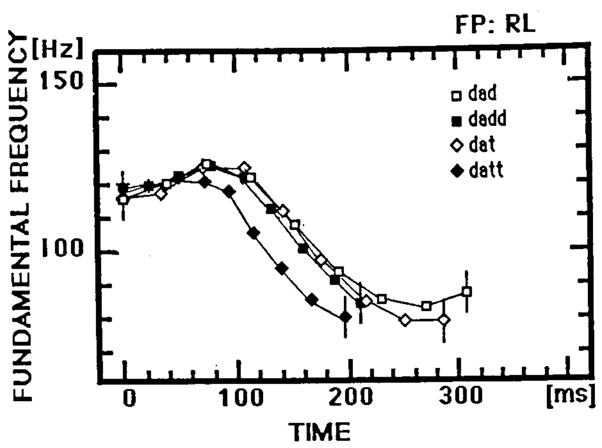


Fig. 3 The fundamental frequency contour of the vowel for different test words in one-word utterances

When focus is assigned to the initial word position the acoustic manifestation of the sentence intonation is in the studied utterances a maximum point of the fundamental frequency contour in the vowel segment of the test word followed by a minimum point, but in this case the minimum point seems to be located outside the vowel segment (fig.4). Most of the frequency fall is, however, still within the vowel segment. Some limited data from another speaker show a somewhat another strategy. For that speaker the fundamental frequency fall is more or less outside the vowel segment. The timing difference of the fundamental frequency fall for the different test words in this position is nevertheless the same as in the other word positions.

The duration of the vowel segment in the different test words is in fig.5. shown for different focus assignments of the three word utterances. The diagram shows that final lengthening is to a great extent dependent on the location of the focus position. When focus is assigned to the initial utterance position the duration of the vowel in the final test word is shortened. The speaker seems to have a prolongation of the vowel segment in the initial word position when it is in focus position that is more or less of the same magnitude as the lengthening process in utterance final position.

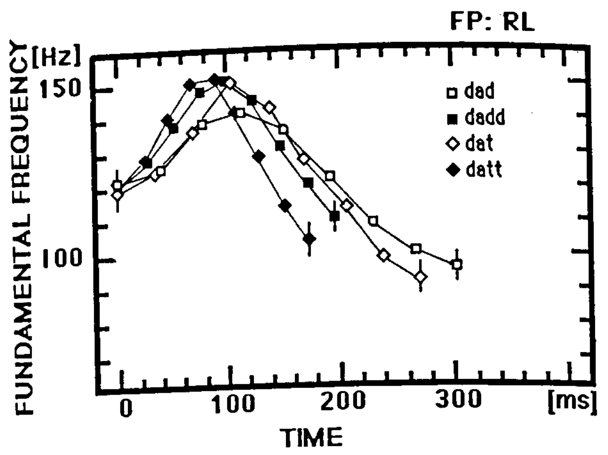


Fig.4 The fundamental frequency contour of the vowel for different test words in initial position of three-word utterances. Focus is assogned to the initial position.

dad

	-foc	+foc
-fin	216	304
+fin	245	330

dadd

	-foc	+foc
-fin	156	198
+fin	162	202

dat

	-foc	+foc
-fin	199	273
+fin	235	294

datt

	-foc	+foc
-fin	132	174
+fin	157	186

Fig. 5 The duration of the vowel segment in msec. is shown for different combinations of focus and utterance positions.

DISCUSSION

The main observation on the timing of the frequency fall connected with the signalling of sentence accent is that the frequency fall will occur earlier in a phonologically short vowel followed by an unvoiced consonant in relation to the CV-boundary than in the other cases. The duration of the vowel is in this case extremely short and the unvoiced consonant cannot convey any information about the fundamental frequency fall. When focus is assigned to the utterance final position and the final word consists of a monosyllable, the frequency fall has to occur within the vowel or within the vowel and the following consonants. When the following consonants are unvoiced, it seems to be necessary to move the frequency fall to an earlier point in relation to the CV-boundary in order not to lose the prosodic information.

It is in non-final position possible to partly locate the fundamental frequency fall to following syllables and words. The speaker in this study locates most of the frequency fall in the vowel but a limited study of another speaker seems to support the idea that it is a possible strategy for some speakers to locate the minimum point of the frequency fall in a following syllable. This is sometimes the case when a monosyllabic word is built up of unvoiced consonants after the vowel. The prosodic information is then signalled by the change of fundamental frequency level in the successive syllables. Similar data can be observed in American English [5].

CONCLUDING REMARKS

It seems possible to assume that an underlying intonation scheme is similar for sentences with the same prosodic pattern but built up of different segments and words. The timing pertubations observed seem to be possible to handle by means of adjustment rules on a more peripheral level. A complete intonation model of a language must at least include the following parts.

- An underlying intonation scheme.
- Timing perturbations owing to the syllable composition.
- Frequency perturbations owing to physiological factors such as e.g. inherent pitch.

The importance of accounting for the different types of perturbations of the fundamental frequency contour in the generation of synthetic speech to obtain a higher degree of naturalness and intelligibility must be determined by perceptual tests.

REFERENCES

[1] Bruce, G. (1977): "Swdish Word Accent in Sentence Perspective". Travaux de l'Institut de Linguistique de Lund XII. B Malmberg and K. Hadding (eds.), Lund : Gleerup.

[2] Pierrehumbert, J. (1980): "The Phonology and Phonetics of English Intonation". Ph.D. Thesis, MIT.

[3] Lyberg, B (1981): "Some observations on the vowel duration and fundamental frequency contour in Swedish". Journal of Phonetics 9, pp. 261-272.

[4] Elert, C.-C. (1964): "Phonological Studies of Quantity in Sweidsh", Almqvist & Wiksell, Uppsala, Sweden.

[5] Lyberg, B. (1984): "Some fundamental frequency perturbations in a sentence context". Journal of Phonetics 12, pp. 307-317.

DURATIONAL PATTERN OF RUSSIAN SYNTAGMA: THE STANDARD SCHEME AND ITS MODIFICATIONS

Olga Krivnova

Department of Philology, Moscow State University
Moscow, USSR, 119899

ABSTRACT

In the present paper it is suggested the idea of the existence of a pronunciational background related to rhythmical speech organization. In temporal aspect this background is realized in particular through some standard scheme of phonosyntagma pattern. On the basis of coherent Russian text the most important qualitative features of the standard scheme are revealed and a numerical model and statistical characteristics of its realization are presented.

INTRODUCTION

The principle of speech rhythmical organization suggests forming a certain pronunciational background in speech signal. Regular reproduction of the elements realizing this background leads to disintegration of speech into various phonetic constituents /phonowords, syntagmas, phrases, periods/ with their specific inner structure periodically recurring in rhythmically organized speech. Rhythm-forming elements having various physical realization attach important functions to acoustic speech parameters - functions of form construction and integrity of speech units. The subject functions are intrinsic to duration as well. In this paper a possibility of presentation of phonosyntagma durational patterns in Russian as a result of realization of a standard scheme and its regular modifications is considered. Besides, the standard scheme is viewed as temporal stereotype which, being a part of pronunciational background and skills of a native speaker, plays an important role in creating phonetic integrity of the syntagma.

STANDARD SCHEME OF SYNTAGMA TEMPORAL ORGANIZATION

Qualitative description

A great number of phonetic experimental research as well as speech synthesis practice demonstrate that the general tendency of syntagma temporal arrangement in Russian is related to forming a positional contrast or, in other words, differences in duration of various components of a phonetic word depending on its position in the syntagma. Various language material shows an almost universal character of the positional factor when analysing its influence on word durational characteristics within syntagma. The positional contrast in a Russian syntagma is clearly detected not only when isolated phrases are pronounced but in running speech as well. Thus we can speak about the existence of a corresponding temporal stereotype or a standard realization scheme of the positional contrast. At the same time up to now the positional contrast has been studied on separate phrases with a limited set of rhythmical word-patterns and sound composition. But if we try to analyse running speech, the data obtained from such a material are not sufficient. In view of the above-mentioned facts the first part of our research was devoted to the qualitative analysis of intrasyntagmatic duration relations in coherent text. Our purpose was first of all to enrich and make more accurate the experimental data obtained earlier.

The study was carried out on a corpus of syntagmas singled out as a result of auditory analysis of coherent scientific Russian text read by an announcer /standard Moscow speaker/ with moderate individual speaking rate.

Syntagmas of various length with a main stressed word in final position were selected for this analysis. Total volume of the sample comprised 438 units. Sound durations were measured according to oscillograms /registered at film speed of I00 mm/sec/. In order to make the segmentation procedure easier speech signal and intensity curve were recorded on the film simultaneously.

As a result of the analysis the following peculiarities of the positional contrast /PC/ were revealed:

I. A normative /statistically predominant/ way of a PC realization is time shortening of a word in the syntagma non-final position. The shortening is achieved by fast speaking rate of these words while maintaining normal speaking rate for words in the final position and single-word syntagmas. The following symbols can be used to designate the way of the PC construction: F /nf/ - N /f/. Further we'll speak about it as of standard temporal scheme /STS/ of the syntagma. Normative nature of the subject scheme is demonstrated by the results of the comparison of the word speaking rate characteristics in various syntagma positions with data on individual speaking rate differences in Russian. Tables I and 2 present data on average sound duration for various individual speaking rates and for words in different syntagma positions accordingly.

Table I. Average sound duration /in msec/ of individual speaking rate in Russian /I/

INDIVIDUAL SPEAKING RATE	FAST	NORMAL	SLOW
GENERAL	65	65-73	73
VARIATION WITHIN NORMAL GENERAL TEMPO	60,4-I,4	74,2-3,0	87,I-2,2

Table 2. Average sound duration /in msec/ for words in different syntagma positions

NUMBER OF WORDS IN A SYNTAGMA	POSITION INITIAL	MEDIAL	FINAL
2	60,0-2,0	---	77,0-3,0
3	59,0-I,4	58,0-I,2	76,0-2,0

Data on word stress perception in a Russian syntagma /2/ make it possible to assume that the scheme F /nf/ - N /f/ being a standard method of the PC realization takes part in the formation of a syntagma accent pattern /component known as syntagmatic stress/. In this connection one may notice that in works on Russian phonetics /3/, /4/ the idea that an increase of the speaking rate can be considered as a source of numerical and qualitative reduction of vowels in a word was put forward more than once. The notion that the fast tempo of pronunciation of syntagma non-final words decreases their prominence level and creates the contrast needed for syntagmatic stress is a natural widening of this idea.

2. Apart from the positional factor, the word duration depends also on such shortening factors as the number of syllables per word and the word distance from the syntagma beginning point. Under the total influence of all shortening factors duration is decreasing in a nonlinear manner showing what is known as "an incompressibility effect" /5/. This effect is clearly seen in syntagma non-final positions /fig. I/.

3. Minimal vowel durations characterizing the incompressibility effect are close to certain temporal perception constants. For example, the minimal duration of an unstressed vowel /$T_{min} \approx$ 30-40 msec/ is close to the threshold value of its detection under any consonant environment /6/. The minimal duration of a stressed vowel /$T_{min} \approx$ 70-90 msec/ is close to threshold value needed for its phoneme running identification /6/. It follows that vowel duration in syntagma non-final words is constrained within limitations which, first of all, provide a possibility of correct identification of rhythmic type of the word and recognition of its stressed vowel. It is also worth mentioning that minimal durations of stressed and unstressed vowels relate as 2 to I.

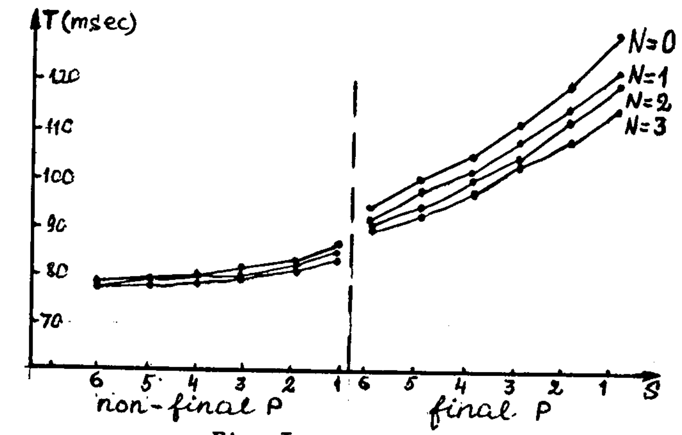


Fig. I

Duration modification of the stressed vowel in syntagma under the influence of the following factors: position /P/, number of syllables per word /S/, word distance from the syntagma beginning point measured in lexical stresses preceded /N/. Average values disregarding vowel quality distinctions.

4. Duration boundary values which separate stressed vowel realizations in non-final and final words /$T_{bnd} \approx$ 90-I00 msec/ are close to phoneme boundary values obtained for languages with phonemic contrast in vowel length /7/, /8/. This makes it possible to speak about different duration categories in Russian speech as well. The analysis shows that a vowel of a mid length is realized when it is stressed, belongs to syntagma final word but not followed by a pause. In that case its duration is about 3 times longer than the critical value of vowel detection. If the latter is classified to be supershort, then vowel realizations of other categories are characterized by the following relations of durations - I/supershort/: 2/short/: 3/mid/: 4/long/.

It is obvious that STS supposes the stressed vowels of non-final words in syntagma to be short and the stressed vowels of final words to be mid.

5. Positional shortening of non-final words in Russian syntagma is of an asymmetric nature: the shortening is mostly noticeable in vowels in the word terminal part beginning with its stressed vowel /"swallowing" of word terminals is an extreme manifestation of this peculiarity/.

Asymmetry leads to smoothing and actually to the loss of temporal contrast of prestressed and stressed vowels in the syntagma non-final word and this apparently hampers correct stress identification.

Numerical model

The second part of our study is devoted to the development of a numerical model of the PC standard temporal scheme. Such a model is of great interest from the various points of view. Generality of the positional factor in its influence on speech temporal characteristics causes us to think that specific linguistic features of the PC are imbodied in its numerical parameters.

The results of the preliminary qualitative analysis give us an idea of the general form of the STS numerical model. While determining concrete numerical estimates we limited ourselves to the consideration of stressed vowels /V́/. Data of their duration were obtained by using the same corpus of syntagmas that has already been described above.

The proposed STS model accounts for all the shortening factors: word position in the syntagma /P/, the number of syllables per word /S/, the number of preceding lexical stresses /N/.

Elongation factors such as absolute prepausal position and position under semantic accent are not taken into account. Consonant environment that influences V́ is not taken into consideration as well. Thus, formulae /I, 2/ give mean duration to the non-prepausal V́ in a syntagma with final main stress when there is no pronounced semantic accent in it.

The general form of the model was chosen on the basis of the notions that were used before in some generative models of speech temporal organization /9/, /I0/.

These notions are as follows:

I. For every stressed vowel of a given phonetic quality there can be singled out two specific realizations: first in the context where both elongation and shortening factors are absent and second in the context where shortening factors have their maximum effect. In the first case it is natural to regard the vowel duration as its intrinsic duration /T_0/, and in the second case as its critical duration /T_{min}/. The difference /$T_0 - T_{min}$/ which characterizes a stressed vowel temporal potential can be called its residual duration.

2. The shortening factors P, S, N effect /$T_0 - T_{min}$/ and besides they act independently and irrespective of vowel quality.

3. The factors S and N shorten the stressed vowel recursively, i.e. their shortening effect is realized cyclicly as a function of the corresponding variables.

In accordance with the accepted assumptions the V́ duration in the absence of elongation factors can be presented as follows:

$$T_{S,P,N}(\acute{V}) = (T_0 - T_{min}) \cdot \alpha^{S-1} \cdot \beta^{P} \cdot \gamma^{N} + T_{min} \quad (1),$$

where S, N as defined above, P can acquire 2 meanings: 0 - for a word in syntagma final position and I - for a word in non-final position, α, β, γ - shortening coefficients.

As a result of the examination of V́ measured durations the α, β, γ -coefficients were approximated as 0,82; 0,20; 0,90 accordingly. In compliance with the obtained estimates the subject suprasegmental factors can be ordered according to their increasing shortening effect as follows: N, S and P. Such an order corresponds to the data presented in other publications.

With the received estimates, the STS model for syntagma stressed vowels looks like this:

$$T_{S,P,N}(\acute{V}) = (T_0 - T_{min}) \cdot 0{,}82^{S-1} \cdot 0{,}20^{P} \cdot 0{,}90^{N} + T_{min} \quad (2).$$

In formula 2 one can find the incompressibility effect detected empirically - vowel reduction decreases gradually when the shortening effect is increased.

Identity of formula 2 was verified on a corpus of 505 syntagmas /the same speaker performed the reading/. As a deviation value of measured and calculated durations the following value was selected: $|T^{msd}_{S,P,N}(\acute{V}) - T^{calc}_{S,P,N}(\acute{V})|$. This value was defined on different generalization levels of durations measured. We have established 3 levels: I - a level of average values of a generalized V́ /disregarding its phonetic quality/: T_0=I30 msec, T_{min}=75 msec, II - a level of mean values of V́ with regard of close-open distinctions: T_0^{a} =I47 msec, T_{min}=85 msec; $T_0^{c,e}$= I34 msec, $T_{min}^{c,e}$ =78 msec; $T_C^{i,u,ɨ}$ =II4 msec, $T_{min}^{i,u,ɨ}$= 66 msec, III - a level of V́ specific realizations: the values T_C and T_{min} are the same as in case II. Sample mean value $|T_{msd.} - T_{calc}|$, deviation and 95% confidence interval for the sample mean ($\pm 2 \frac{\sigma}{\sqrt{n}}$) were calculated for each generalization level. Table 3 shows the calculation results.

Table 3. Statistical characteristics of $|T_{msd} - T_{calc}|$ for different generalization levels of empirical data

STATISTICAL CHARACTERISTICS	GENERALIZATION LEVELS I	II	III
THE SAMPLE MEAN /msec/	5,6	7,4	I5,8
MEAN-SQUARE DEVIATION σ /msec/	7,2	7,2	I3,I
$\pm 2 \frac{\sigma}{\sqrt{n}}$	2,2	I,4	0,8
n - NUMBER OF THE ITEMS	4I	96	I03I

Accounting that the accuracy of duration measurement didn't exceed 5 msec the agreement between the calculated and empirical data can be considered as good. Cases of essential deviations demand a close analysis: they give evidence of additional factors the effects of which are not taken into consideration in the proposed model.

STS realization in running speech

The syntagma STS F /nf/ - N /f/ is a component of a pronunciational background stipulated by speech rhythmical organization. It is obvious that a concrete realization of the scheme in a coherent text depends on various contextual conditions. Formula 2 /when used to calculate vowel temporal scales/ shows variation of absolute values of a stressed vowel duration in syntagmas that conform to the STS but at the same time have different verbal filling. It is also important to detect cases of STS disagreement and to reveal their sources. It is worth remembering that violation of general rhythm-forming tendencies is the most relevant informative and descriptive mean in coherent text. Thus the task of detecting the syntagmas corresponding to the STS and those having deviations from it arises.

In this research the same material which served as the basis for the STS numerical model construction was used for the elaboration of the given task. Analysis-by-synthesis procedure was utilized to solve the problem of whether a syntagma chosen arbitraly from the text corresponds to the STS. The latter was considered to be realized in the syntagma if its stressed vowels had the durations compatible with the values calculated by formula 2. Thus empirical data were interpreted taking into account specific phonetic conditions /vowel quality, word rhythmical type, syntagma length/ under which STS is realized. If there was incompatibility of calculated and empirical V́ durations in the syntagma we considered that there was an STS violation in it.

The most difficult problem which arose in the process of implementation of above-mentioned procedure is the definition of critical value sufficient to acknowledge essential divergence of empirical and calculated durations. The difficulty of the problem lies in the fact that very little is yet known about how a listener perceives, estimates and interprets duration differences. We decided to consider a divergence unessential if the value $\frac{|T_{msd} - T_{calc}|}{T_{calc}}$ didn't exceed 20%. It is worth noticing that despite some divergences the overall data at present show that the critical value of just-noticeable difference for the perceived change in vowel duration is about I5%-20% /9/.

Before presenting the results of the experimental research conducted it is worth showing theoretically possible types of deviations from the STS. Since the STS presupposes the realization of the stressed vowel in the syntagma non-final and final words in short and mid lengths accordingly, it is easy to see that at the level of stressed vowels the following deviations from the standard are possible:

A - absence of V́ reduction in the syntagma non-final word /i.e. realization of the mid or long category/,

B - V́ reduction in the syntagma final word /i.e. realization of the short category/,

C - V́ lengthening in the syntagma final word /i.e. realization of the long category/.

Table 5 displays statistical data /%/ on STS realization obtained on the basis of our research material.

Let's look at the section I first. We can see that when the main stressed word is in final position the STS is realized without any deviations in approximately half of the cases. Within the set of syntagmas with STS violations those with one deviation prevail, syntagmas with more than one deviation constitute only II% of the overall set. It follows that in syntagmas of this type the observed deviations are local modifications of the standard and are not the result of its modification as a whole. The same conclusions may be deduced for syntagmas settled in section II.

Table 5. STS realization in coherent text
I - syntagmas with main stressed word in final position /438 units/, II - syntagmas with main stressed word in non-final position /67 units/.

	NUMBER AND TYPES OF DEVIATIONS FROM THE STS 0	I			2			3
	0	A	B	C	A,A	A,B	A,C	A,C
I	45,9	9,6	I9,4	I4,8	0,9	6,6	2,3	0,5
II	28,4	I9,4	I9,4	3,0	3,0	I4,9	9,0	3,0

On the basis of literary data and the results of our research the following sources and positions of probable STS disagreements can be distinguished:

I. Absolute prepausal V́ position/for final word/.
2. Type of pitch contour /for word in main stressed position/.
3. Weakening or strengthening of a word prominence level /for word in any position/.
4. Syntagma position in regard to utterance external and internal boundaries /for word in final position and position of main stress/.

REFERENCES

/I/ Агафонова Л.С., Бондарко Л.В. и др. О некоторых характеристиках русской речи в зависимости от разных темпов произнесения. - В кн.: Слух и речь в норме и патологии. Л., 1974, I, с. I0-I7.
/2/ Светозарова Н.Д. Интонационная система русского языка. Л., 1982.
/3/ Чистович Л.А. и др. Речь. Артикуляция и восприятие. М.-Л., 1965.
/4/ Бондарко Л.В. Фонетическое описание языка и фонологическое описание речи. Л., 1981.
/5/ Klatt D.H. Interaction between two factors that influence vowel duration. JASA, I983, v.54, p. II02-II04.
/6/ Чистович Л.А. и др. Восприятие речи человеком. Л., 1976.
/7/ Fujisaki H., Nakamura K., Imoto T. Auditory perception of duration of speech and non-speech stimuli. Ann. Bull. of Res. Inst. of LPhUT, I973, 7, p. 45-64.
/8/ Nooteboom S.G. Contextual variation and perception of phonemic vowel length. - Preprints of the Sp. Com. Sem., Stokholm, I974, 3, p.I49-3.
/9/ Klatt D.H. Linguistic uses of segmental duration in English:acoustic and perceptual evidence. JASA, I976, v. 59, p. I208-2I.
/I0/ Lindblom B., Lyberg B., Holmgrem K. Durational patterns of Swedish phonology: do they reflect short-term memory processes? Indiana Univ. Ling. Club, I98I.

SYSTEMATIQUE DES DUREES SEGMENTALES DANS LES RIMES SYLLABIQUES

A VOYELLES LONGUES ET BREVES PAR NATURE.

LAURENT SANTERRE

Département de Linguistique

Université de Montréal

C.P. 6128, Montréal, H3C 3J7. Canada

RESUME

Le français québécois comporte des voyelles phonologiquement longues ou brèves, de même que des consonnes phonétiquement longues ou brèves, dont certaines sont abrégeantes et d'autres allongeantes. Les rimes syllabiques qui en résultent présentent des configurations systématiques de durées qui tiennent à la nature des noyaux et des codas. Cette systématique peut être exploitée pour la synthèse et la reconnaissance automatique de la parole naturelle.

INTRODUCTION

Le français québécois a conservé l'ancien système phonologique de 17 voyelles, dont huit comportent nécessairement, en plus de la distinction du timbre, le trait de durée en syllabe entravée. Ces voyelles longues sont le /3/ de fête opposé au /E/ bref de faite, le /A/ de pâte opposé au /a/ bref de patte, le /O/ de côte opposé au /o/ bref de cote, le /eu/ de jeûne opposé au /oe/ bref de jeune, et les quatre nasales /en/ de feinte, /an/ de fente, /on/ de fonte et /un/ de défunte.

Dans le but de programmer la synthèse par règles du français du Québec de niveau international, on a entrepris d'étudier la prosodie très mal connue de ce dialecte, et en particulier la systématique des durées segmentales en fonction des positions accentuelles. Les résultats que nous présentons ici portent sur un corpus de 215 mots prononcés isolément puis repris en fin de syntagme sujet; ex."pâte"; "le mot pâte me plaît". Ce corpus a été enregistré plus d'une fois, toujours dans les mêmes conditions et à débit constant par la même personne,la production d'un ou deux autres locuteurs québécois ne servant que de contrôle. Les résultats présentés ici font donc ressortir la systématique des durées chez ce seul locuteur; car si une telle systématique existe, elle doit s'exercer et apparaître plus nettement que si on l'extrait des moyennes de plusieurs locuteurs. L'ordinateur n'a pas à parler comme une moyenne, et il n'aura jamais à reconnaître la parole d'une moyenne de locuteurs. La systématique de chaque locuteur respecte celle de son dialecte mais reste personnelle. Dans ce premier temps, c'est cette systématique des durées que nous avons voulu dégager.

Nous n'avons retenu pour l'instant que les rimes syllabiques comportant les consonnes dites obstruantes, à l'exclusion des sonantes; il s'agit donc des 7 voyelles brèves et des 8 voyelles longues entravées par 12 consonnes, soit les 3 occlusives sourdes, les sonores correspondantes, les 3 constrictives sourdes et les sonores correspondantes. Nous n'avons pas encore tenu compte de la légère influence de la consonne initiale de syllabe, ni des durées spécifiques des voyelles qui sont bien connues (Di Christo 1980). Ces durées fines seront prises en compte au moment de la formulation des règles de base de chaque phonème.

Groupes de voyelles et de consonnes

L'examen des durées vocaliques nous permet d'établir des groupes homogènes de voyelles selon leurs comportements sous l'entrave abrégeante ou allongeante; ce sont les voyelles hautes i, y, u; les quatre brèves E, a, o, oe; les quatres longues correspondantes 3, A, O, eu; les quatre nasales. Les deux voyelles restantes,/e/ et le schwa, ne se rencontrent pas en syllabe entravée.

Les consonnes se divisent nettement en fortes ou longues, et en faibles ou brèves, selon qu'elles sont sourdes ou sonores. Les consonnes longues par nature sont P T K et surtout f s C (comme dans chat); on verra par les tableaux qui suivent que seules les occlusions sourdes sont abrégeantes. Les consonnes brèves par nature sont les occlusives sonores et les constrictives sonores, ces dernières seules étant allongeantes. Les abrègements et les allongements sont très prononcés sur les sept voyelles brèves par nature, et beaucoup moins sensibles, bien que réels, sur les huit voyelles longues par nature.

Dans les tableaux qui suivent, on trouve dans l'ordre la durée de la voyelle, celle de la consonne, la durée de la rime et l'écart des pourcentages de durée occupée par les deux segments. Cet écart est marqué positif si le noyau vocalique domine, et négatif si la coda l'emporte en durée. Ex. i y u + P T K: la voyelle occupe 30.2% d'une rime de 26.6 cs, la consonne 69.8%; l'écart négatif est de -39.6.

Le tableau 1 présente les résultats sous l'accent en finale abso lue, le tableau 2, sous l'accent en fin de syntagme intérieur.

Tableau 1

		V.	C.	R.	E.
1.	i y u + P T K	8.	18.56	26.6	-39.6
	E a o oe -	10.	17.	27.	-26.
2.	i y u + b d g	11.1	11.95	23.	- 3.6
	E a o oe -	13.88	11.84	25.72	+ 8.
3.	i y u + f s C	11.64	23.12	34.76	-33.
	E a o oe -	12.21	21.35	33.56	-28.
4.	i y u + v z J	24.16	9.3	33.46	+44.4
	E a o oe -	19.84	10.6	30.44	+30.
5.	3 A O eu + P T K	19.34	16.83	36.17	+ 7.
	an en on un -	20.61	13.12	33.73	+22.
6.	3 A O eu + b d g	20.58	9.87	30.45	+35.
	an en on un -	23.18	8.95	31.29	+39.5
7.	3 A O eu + f s C	21.72	18.7	40.42	+ 7.4
	an en on un -	24.47	17.62	42.	+16.
8.	3 A O eu + v z J	24.31	10.85	35.16	+38.
	an en on un -	28.75	9.83	35.58	+49.

Moyennes de durées sous l'accent en finale absolue.

Tableau 2.

		V.	C.	R.	E.
1.	i y u + P T K	5.26	13.28	18.54	-43.4
	E a o oe -	8.18	14.	22.18	-26.4
2.	i y u + b d g	7.25	6.25	13.75	+ 5.4
	E a o oe -	10.53	7.58	18.11	+16.
3.	i y u + f s C	7.18	13.26	20.45	-29.6
	E a o oe -	9.0	12.32	21.32	-15.4
4.	i y u + v z J	14.65	7.94	22.59	+29.7
	E a o oe -	13.32	8.26	21.58	+24.
5.	3 A O eu + P T K	13.85	11.73	25.58	+ 8.
	an en on un -	15.57	11.38	26.26	+15.5
6.	3 A O eu + b d g	15.48	7.35	22.83	+36.
	an en on un -	16.87	8.15	25.0	+35
7.	3 A O eu + f s C	15.37	11.3	26.67	+15.3
	an en on un -	17.96	13.87	31.83	+13.
8.	3 A O eu + v z J	17.19	7.9	25.09	+37.
	an en on un -	19.41	8.7	28.11	+38.

Moyennes des durées sous l'accent de fin de syntagme intérieur.

Durées des groupes vocaliques sous l'accent terminal.

Les brèves par nature (non allongées par v z J (comme dans Jean)) ont des durées qui se situent entre 8 et 14 cs, avec une moyenne à 11.64 et un sigma de 2. Par comparaison dans le groupe on peut considérer que P T K les abrègent un peu (8 et 10 cs) que b d g les allongent un peu (11.1 et 13.88 cs), tandis que f s C les laissent inchangées (11.64 et 12.21 cs).

Les voyelles orales longues par nature affichent des durées moyennes de 21.48 cs entre 19 et 25 cs et un sigma de 2.1. Elles sont un peu abrégées par P T K (19.34), très peu par b d g (20.58); suivies de f s C, elles restent près de la moyenne (21.72), et devant v z J elles s'allongent à 24.31 cs. Les voyelles brèves par nature allongées par v z J sont presque aussi longues que les longues par nature, elles-mêmes allongées: (24.16 pour les voyelles hautes, et 19.34 pour les autres brèves). On peut remarquer que ces dernières se démarquent des longues par nature, par exemple dans des paires comme lève et rêve, page et l'âge, loge et l'auge.

Enfin, les nasales sont plus longues que toutes les autres voyelles dans les mêmes conditions. Leur moyenne se situe à 24.25 cs entre 20 et 29, avec un sigma de 3.4. Elles sont un peu abrégées par b d g (23.18), laissées inchangées près de la moyenne par f s C (24.47) et allongées par v z J (28.75).

Durées des groupes vocaliques sous l'accent intersyntagmatique

Sous l'effet de la désaccentuation qu'entraîne le déplacement de la syllabe accentuée sur la fin du sujet, les durées subissent des diminutions systématiques. Les voyelles brèves ont une moyenne de 7.9 cs (contre 11.64 sous l'accent terminal) située entre 5 et 11 cs (avec un sigma de 1.8). Les voyelles orales longues par nature voient leur moyenne réduite à 15 cs (contre 21.48) entre des durées extrêmes de 13 et 17.5 cs avec un sigma de 1.37. Les nasales se situent en moyenne à 17.5 cs (contre 24.25 en finale absolue) entre 15 et 19.5 cs avec un sigma de 1.5. Le tableau 3 rappelle ces comparaisons et montre les pourcentages de réduction due à la position accentuelle.

Tableau 3

	Accent terminal		Accent intérieur		
	M	S	M	S	%
Brèves	11.3	2	7.9	1.8	71
Longues	21.48	2.1	15.	1.37	70
Nasales	24.25	3.4	17.5	1.5	72

On peut voir que l'effet de la désaccentuation sur la durée a été sensiblement le même pour les trois groupes de voyelles.

Durées des groupes de consonnes sous l'accent terminal

Ce qui est à montrer ici, c'est que certains groupes de consonnes règlent leur durée à l'intérieur de la rime sur la durée vocalique; ainsi, en reprenant les données du tableau 1, on voit que les durées des P T K diminuent à mesure que les voyelles entravées s'allongent:

Tableau 4

	V	C	R	E
i y u + P T K	8.03	18.56	26.6	-40
E a o oe -	10.	17.	27.	-26
3 A O eu -	19.34	16.83	36.17	+ 7
an en on un -	20.61	13.12	33.73	+22

Durée des groupes consonantiques

Ce phénomène est plus prononcé encore après les nasales où les consonnes ne font plus que 13.12 parce que les voyelles très longues par nature imposent leur prépondérance dans la rime. Tout se passe dans l'organisation des durées relatives comme si, tout en respectant les exigences de durée longue ou brève des voyelles et aussi celles des consonnes longues (sourdes) ou brèves (sonores) par nature, la durée de la rime constituait un cadre quantitatif limitatif. Avec les voyelles brèves par exemple, la rime se situe autour de 27 cs et ce sont les durées des segments qui composent entre eux pour respecter cet ordre de grandeur; c'est ce qui fait varier l'écart des pourcentages de durées relatives.

Avec les deux groupes de voyelles longues par nature, la rime est nettement plus longue (33-36 cs), mais cet ordre de grandeur est relativement respecté. Voyons s'il en est de même avec les autres consonnes.

Avec les b d g, brèves par nature, la durée de la rime à noyau vocalique bref est de 23-25.72 cs (voir tableau 1), tandis qu'elle est de 31 cs environ quand le noyau vocalique est une longue par nature. La durée des consonnes diminue quand celles des voyelles augmentent.

Avec les longues par nature f s C, encore plus longues que les P T K, les rimes sont de l'ordre de 34 cs quand le noyau est bref, et de 40-42 cs quand il est long. On observe le même jeu de compensation voyelle-consonne: la durée consonantique varie en sens inverse par rapport à la durée vocalique. Enfin les consonnes allongeantes et brèves par nature v z J obéissent au même mécanisme; mais cette fois le noyau vocalique est toujours long, de sorte que la rime n'a qu'un ordre de grandeur: de 30 à 38.5 cs. Mais on ne voit pas les consonnes brèves descendre sous

le seuil des 9 cs, même quand les voyelles s'allongent considérablement comme les nasales. Il en résulte des écarts positifs (en faveur du noyau vocalique) très marqués dans ce groupe de rimes (44 et 49).

On peut répartir les rimes syllabiques du point de vue des Ecarts en deux groupes, négatif et positif, qui présentent une interface. Les rimes comprises dans cette interface peuvent être nettement séparées par les constituants ou par la durée de la rime elle-même.

On peut entrevoir, même par cet exemple limité, qu'un programme d'intelligence artificielle appliqué à la systématique des durées à l'intérieur de la rime syllabique sous cet accent peut grandement contribuer à la reconnaissance automatique des segments. L'espace manque pour faire le même examen sur les données pour l'accent intérieur; on peut le faire au moyen du tableau 2. Nous nous contenterons de voir l'effet global de la désaccentuation sur la durée des consonnes.

tableau 5

	Voyelles brèves			Voyelles longues		
	acc.final	acc.inter	%	acc.final	acc.inter	%
P T K	17.78	13.64	76.7	1.5	11.55	77.
b d g	11.89	7.	58.8	9.41	7.75	82.3
f s C	22.23	12.79	57.5	18.16	12.58	69.3
v z J	10.	8.1	81.	10.34	8.3	80.

Réduction des consonnes par désaccentuation

On peut voir que les abrégeantes et les allongeantes se réduisent respectivement à 77 et 80% environ, et cela aussi bien après les voyelles brèves qu'après les voyelles longues par nature. Pour les constrictives, cela peut se comprendre, puisque les brèves sont presque aussi longues que les autres sous leur entrave. D'ailleurs elles sont déjà assez brèves et ne sauraient se réduire beaucoup plus. Les 3 occlusives abrégeantes doivent garder leur trait de longueur et leur effet abrégeant avec toutes les catégories de voyelles.

Ces considérations pourraient sans doute servir à établir des règles de durée très réalistes pour la synthèse de la parole, mais paraissent plus difficiles à exploiter pour la reconnaissance que sous l'accent terminal.

Enfin, on peut remarquer que les différences de durée entre les deux groupes d'occlusives et entre les deux groupes de constrictives est un moyen plus sûr de distinguer les cognates que le trait dit de sonorité. Les occlusives diffèrent par 35% de durée sous l'accent terminal et par 32% sous l'accent intérieur; les constrictives sonores sont deux fois plus courtes que leurs cognates sous l'accent terminal, et plus courtes de 35% sous l'accent intérieur.

Conclusion

Cette étude sur les durées fait partie d'une recherche plus vaste sur la prosodie du français québécois, en vue de la synthèse et de la reconnaissance automatiques. Les durées relatives dans les rimes syllabiques répondent à une systématique bien définie qui repose, dans ce dialecte, sur un système de voyelles brèves et de voyelles longues par nature, et sur des consonnes qui les abrègent toutes et d'autres consonnes qui les allongent toutes, i.e. les occlusives sourdes et les constrictives sonores. Les sourdes sont elles-mêmes longues par rapport aux sonores. Les occlusives sonores et les constrictives sourdes n'ont que très peu d'influence sur la durée des voyelles qu'elles entravent.

Par contre, dans toutes les rimes longues ou brèves, il s'exerce un jeu de prédominance systématique de la durée vocalique ou consonantique, selon la nature des segments en présence. Cette systématique rend possible l'utilisation de l'intelligence artificielle dans la programmation de la synthèse et de la reconnaissance de la parole naturelle.

BIBLIOGRAPHIE

-Di Christo,A. (1980), "La durée intrinsèque des voyelles du français". Travaux de l'Institut de Phonétique, Aix, vol. 7, p. 211-235.

-Jacques, B. (1973), "Variations de durée des voyelles et des consonnes frivatives post-vocaliques finales en position accentuée et inaccentuée", Manuscrit, Université du Québec, Canada.

-O'Shaughnessy, D. (1981), "A Study of French (Canadian) Vowell and consonnant Durativity", Journal of Phonetics, p.385-406.

-Santerre, L. (1974), "Deux E et deux A phonologiques en français québécois". Le français de la région de Montréal", les Cahiers de linguistique, no. 4, Presses de l'Université du Québec, Canada, p.117-145.

PROSODIC INTERFERENCE: A TYPOLOGICAL APPROACH

ANNA A. METLYUK

Minsk State Pedagogical Institute of Foreign Languages
Minsk, Byelorussia, USSR, 220662

ABSTRACT

The interaction of prosodic systems in a bilingual's speech reveals itself in minimal prosodic units (tonemes, accentemes, chronemes, rhythmemes) and in their structural complexes, or "phonological syntagms" (tonal contours, accentual, temporal and rhythmic structures). As the actual relations between the units of the languages in contact are set by a bilingual speaker according to the laws of interlanguage identification, the character of these relations determines types of interference on the paradigmatic plane (underdifferentiation, overdifferentiation, substitution) and on the syntagmatic plane (intercatenation, plus-segmentation, minus-segmentation, permutation).

INTRODUCTION

Prosodic interference is defined as changes in the realization of the prosodic system of the non-native, second language (L_2) that emerge under the influence of the native language (L_I), and manifest themselves in a bilingual's speech as deviations from the norm of L_2.
Topicality of problems of prosodic interference for language theory and applied linguistics has widened the range of experimental phonetic investigations in spite of the lack of knowledge on prosody as linguistic phenomenon and despite the difficulties of contrastive prosodic analysis aimed at revealing areas of potential interference in a bilingual's speech. Investigations are carried out predominantly on the level of the utterance (intonation group) in terms of perceptual and acoustic features, pertaining to the prosodic structure of an utterance as a whole and to its separate elements - prehead, head, nucleus and tail.
The features of prosodic interference (deviations, errors) are analysed as to their frequency, stability, communicative relevance as well as to their occurrence in different types of utterances and prosodic subsystems.
The typology of prosodic interference bases on the character of actual relations between the elements of the two languages as they are set by a bilingual speaker according to the laws of interlanguage identification has not as yet been touched upon. The solution of this significant linguistic problem requires the description and classification of the prosodic units of the languages in contact. But what units does the prosodic system comprise? The question, as we know, is a point of discord between linguists and an attempt is made here to give our interpretation of the units of a prosodic system.

TYPES OF PROSODIC UNITS

In view of the polycomponental and polyfunctional nature of prosody it seems logical to admit objective existence of essentially heterogenious prosodic units, forming relatively autonomous but interconnected and interpenetrating subsystems - temporal, accentual, rhythmical and tonal. Each of the subsystems contains units of two types - microprosodemes (minimal prosodic units) and macroprosodemes (structural complexes of microprosodemes, "phonological syntagms").The first type includes syllablechronemes, accentemes, rhythmemes,tonemes (functional types of syllable duration,accents, rhythmic units, tones). The second type is represented by temporal, accentual, rhythmic and tonal structures which function as integrative units, as patterns of syntagmatic (phonotactic) organization of microprosodemes.
Each structural complex is at the same time a paradigmatic unit when opposed to other structural complexes of the subsystem.
The units of both types are phonological units, if phonology is viewed in the broad sense including(i)segmental and suprasegmental units,(ii) elementary units and "phonological syntagms",(iii)units that perform a distinctive function and those that have no such function in the language but fulfil the constitutive (integrative) and identificatory functions. In the functioning of prosody as one whole the structural correlation of the toneme, accenteme and chroneme as systemic elements provides their close interconnection and interaction in complex polycomponental units - micro- and macroprosodemes of an utterance.
Ontologically, prosodemes as invariants exist in classes of their variants (alloprosodemes) as the general in the particular. And since the prosodeme, unlike the phoneme, is a sign, the invariability of the prosodeme is inherent in both its aspects - form and content. The invariant of the prosodemic form is its material essence (its constant phonetic features) and the invariant of the content is a generalized denotational meaning. Thus, the rising toneme has a rising direction of pitch movement as its invariable formal feature and its denotational (logical-modal, or intellective) meaning of indefiniteness, non-finality,incompleteness is its semantic invariant, which in its turn is conveyed, irrespective of the context, by all the functional and structural variants of the toneme and is realized as categorial meaning of the communicative type of an utterance.
Variants of the toneme as to its form are marked by configurational and pitch-level varieties of its allotonemes (tone types); whereas its variants as to the content are distinguished by subjective-modal (emotive,attitudinal) connotations realized as different situational (or stylistic) and pragmatic meanings of an utterance.
The number of prosodemic variants is conditioned by the system which regulates their (i) positional, combinatory and structural distribution within a polycomponental macroprosodeme (phonological syntagm), the variants being closely interconnected with the phonemic structure of the speech segment, and (ii) semantic-functional distribution, determined by the polysemantic nature of prosodemes on the one hand, and by their interaction with the units of lexical and grammatical subsystems of the language, on the other. The system specifies the areas of realization of prosodemes and limits their variation. On the whole the prosodic system, as any other system, stipulates the norms of its functioning.
In spite of the fact that prosody is to a great extent universal in its categories, languages differ one from another in the number and character of their prosodemes, in the frequency and sphere of their functioning, as well as in the number, distribution and acoustic-perceptual peculiarities of their variants. All that provides the basis for interference, i.e. variations in the form and character of functioning of prosodemes which do not conform to the norms of the second language. In the study of Byelorussian-English, Byelorussian-Russian, Russian-English and other types of interference /I,2,3,4,5/ contrastive experimental analyses of prosodic norms of the languages in contact that preceded error analysis made it possible to establish the inventory of units of the tonal, accentual, temporal and rhythmic subsystems of the languages and as a result to approach the description of prosodic interference in terms of systemic units. An attempt was made to determine types of prosodic interference on the basis of peculiarities of the influence of L_I upon L_2 in the sphere of paradigmatic and syntagmatic realtions between prosodic units /1/.

TYPES OF PROSODIC INTERFERENCE

The interaction of prosodic systems in a bilingual's speech takes place on the level of minimal prosodic units and on the level of phonological syntagms. In the latter cases the interference of the native language system in the syntagmatic organization of the microprosodemes within the structures is more expressed than in the features pertaining to the structure as a whole /1,2/.
Types of paradigmatic interference, as is well known, are determined by the differences in the number and character of prosodic units, by their semantic-and-functional differences in the languages under consideration.
The unequal number of prosodemes in the two languages provides a basis for underdifferentiation or overdifferentiation of some of them by a bilingual speaker.Thus, in the English speech of Russians there occurs underdifferentiation of English rising and falling-rising tonemes, the latter being absent in the system of the Russian language. But in the Russian speech of Englishmen there often appears overdifferentiation of the Russian rising toneme: the structural variants ⌣ of its level-rising ⌣ allotoneme is identified by Englishmen as the falling-rising toneme. It should be noted,however, that due to the universal nature of the majority of prosodic units the above-mentioned types of paradigmatic interference are not frequent.
A more widespread type of paradigmatic interference is substitution. It is a result of language distinctions in the character of prosodemes and the sphere of their functioning, as well as distinctions in the frequency of occurrence of the alloprosodemes that represent them.
The following deviations in the English

speech of Byelorussians and Russians illustrate substitution in all the subsystems of prosody: 1) the use of the rising toneme in non-final intonation groups instead of the falling toneme and vice-versa, the use of the falling toneme instead of the rising one in requests, apologies, contradictions; 2) the substitution of the English rising toneme for the corresponding Russian toneme in particular, the use of the rising-falling allotoneme ʌ which is more frequent in Russian instead of the more frequent English level-rising allotoneme _/; 3) the use of the Byelorussian rising toneme (its level-rising allotoneme in particular) instead of the English falling-rising toneme in general questions; 4) the use of strong accentemes instead of weak ones due to a greater number of strong (full) accents in Russian and Byelorussian utterances as there is a greater number of notion words in them and some classes of words (personal and possessive pronouns, modal verbs etc.) attract accent more often than in English and the role of the semantic and grammatical factors of accentuation is greater, whereas in English the rhythmic factor is the main regulator of accents; 5) substitution of accentual and temporal structures of rhythmic units, tonal contour substitution; 6) substitutions of the accentemes (nuclear and non-nuclear) of L_2 for the corresponding accentemes of L_1, substitution of syllable chronemes, the use of configurational and pitch-level allotonemes of L_1 instead of the functionally similar allotonemes of L_2.
Underdifferentiation and overdifferentiation, as well as substitution of functionally different prosodic units (points 1-5) belong to communicatively relevant semantic interference, i.e. display both formal and semantic-functional non-standard variation of prosodemes and alloprosodemes. Within communicatively relevant prosodic interference two subtypes are distinguished: modal (emotive) and stylistic.
Substitution of units, which are functionally identical but qualitatively different (point 6) belongs to communicatively irrelevant interference; such substitution signals formal-structural non-standard variation only.
The above-mentioned types of interference are deviations from the norms of prosodic unit choice which occur when the communicative-pragmatic message is actualized in various speech situations.
Syntagmatic interference is represented by the deviations in the combinability of microprosodemes within structural complexes (macroprosodemes), on the one hand, and by the inadequate realization of microprosodemes as certain positional and combinatory variants, on the other.
One type of syntagmatic interference is intercatenation of microprosodemes of L_2 according to the structural pattern of L_1, resulting in a new structural pattern, which does not exist in L_2, e.g. inadequate combination of the sliding (or heterogenious) head with the high rising toneme in the contour of English general questions, pronounced by Byelorussian speakers of English.
Another type is minus-segmentation, or elision of elements in the structure, e.g. the omission of prehead or tail in English utterances of Byelorussian-English bilinguals due to an increased number of accents or shift of accents. Wherefore plus-segmentation, or increase of elements in the structure as, for instance, in cases when the prehead and tail appear in the structure of an utterance is due to the reduced number of accents. The three types are deviations from the structural norms of the functioning system and belong to communicatively relevant interference.
Closely connected with these types are deviations from the norms of realization of L_2 units. The latter are conditioned by language distinctions in the acoustic areas of the units.
Thus realization of syllablechronemes of L_2 which are functionally identical to L_1 chronemes differ in utterances of Byelorussian-Russian (B-R) and Byelorussian-English (B-E) bilinguals. Initial unaccented syllables and the 1st fully accented syllable in B-E and R-E are longer than in standard English realizations (SE), the tempo of pronunciation being identical. The 2nd accented syllable and the unaccented syllables of the head are characterised by approximate isochrony in B-E, whereas the nuclear and post-nuclear syllables are drawled in comparison with the temporal standard of Russian (SR) and English (SE). All that leads to inadequate realizations of temporal structures of L_2.
Distortions in the norms of realization are well marked in R-E and B-E utterances of identical tonal contour. F_0 intervals between the elements of the contour are not as evident as in SE. Prehead and head (its 1st accented syllable) have lower F_0 level in R-E and B-E than in SE. F_0 interval of the gradually descending head is wider in R-E and B-E due to the lowering of F_0 level on the 2nd accentual unit. Interference in the realization of English falling toneme is marked in B-E by the lower initial F_0 level and higher final F_0 level than in SE. Realizations of level-rising allotonemes in B-E are characterised by their higher initial and lower final F_0 levels and, consequently, by a narrower F_0 interval than in SE.
In the realization of utterance accentemes in B-E and R-E /2,3/ syllable prominence is achieved by different combinations of acoustic parameters. The acoustic contrasts of unaccented and accented syllables are less clearly marked in R-E, and especially in B-E realizations as compared to ES. The distortions of the norm as to differences in the distribution of phonetic features of prosodic units are termed permutational interference.
The study of prosodic interference in various types of natural and "classroom" bilingualism /5/ reveals typologically common features of interference, which are characteristic of kindred and structurally similar languages and specific features, characteristic of the speakers of only one language.
Some of the typologically common features are as follows: (i) a higher final F_0 level of the falling toneme in the English and German speech of Russians and Byelorussians and, consequently, a lower final F_0 level of the falling toneme in Russian spoken by Englishmen and Germans; (ii) a lower F_0 level of the tonal contour of English and German utterances in R-E, B-E, R-G and B-G as compared to SE and SG realizations; (iii) drawled initial syllables (accented and unaccented) in B-E and B-G utterances as compared to their standard realizations.
Specific features of interference in particular types of bilingualism are (i) greater I contrasts between accented and unaccented syllables in R-E as compared to B-E; (ii) the absence of reduction in unaccented syllables in B-E; (iii) rhythm in B-E which tends to be syllable-timed.
In the cases of reverse language domination in the same types of bilingualism prosodic changes have an opposite direction.
There should also be mentioned universal deviations such as slowing down of the tempo of utterance, increase in the number of accents, rhythmic distortions etc. which can be observed at the early stages of any type of bilingualism.

CONCLUSION

Description of prosodic interference as a linguistic phenomenon in terms of systemic units of prosody extends our general knowledge of phonological interference /6/ and makes it possible to model the so-called "interlingua" at different stages of bilingualism which is important for the theory of language contacts and for practical application in teaching a second language.

REFERENCES

/I/ Метлюк А.А. Взаимодействие просодических систем в речи билингва.-Минск, 1986.
/2/ Метлюк А.А.,Карневская Е.Б. Некоторые аспекты просодической интерференции. - В кн.: Экспериментальная фонетика. - Минск, 1974. Карневская Е.Б., Метлюк А.А. Определение степени просодической интерференции. - В кн.: Экспериментальная фонетика. - Минск, 1976.
/3/ Поплавская Т.В. Просодия английских восклицаний в условиях интерференции. Дис....канд. филол.наук. - Минск, 1978.
/4/ Метлюк А.А. Из исследований просодической интерференции при искусственном белорусско-английском двуязычии. - В кн.: Лингвистическая интерпретация результатов экспериментально-фонетических исследований речевого текста. Тезисы докладов республиканского симпозиума. - Минск, 1977.
/5/ Метлюк А.А.,Евчик Н.С.,Карневская Е.Б. и др. Просодическая интерференция в иноязычной речи. - Минск, 1985. Зарецкая Е.В, Лавенкова Л.А.,Петрушенко Е.Т. О немецко-русской интерференции на просодическом уровне.- В кн.: Лингвистическая интерпретация результатов экспериментально-фонетических исследований речевого текста. Тезисы докладов республиканского симпозиума. - Минск, 1977. Блохина Л.П. Интонационный аспект билингвизма (на материале немецкого и русского языков). - В кн.: Актуальные вопросы интонации.-Москва,1984.
/6/ Weinreich U. Languages in Contact. Findings and Problems. - Hague and Paris, Mouton, 1970.

PHONETIC INTERFERENCE IN BILINGUALS' LEARNING OF A THIRD LANGUAGE

JOAQUIM LLISTERRI

DOLORS POCH-OLIVÉ

Laboratori de Fonètica, Facultat de Lletres,
Universitat Autònoma de Barcelona, Bellaterra, Barcelona,
Spain.

ABSTRACT

This paper presents three experiments concerning the acquisition of French and English as L2 or L3 by bilingual and monolingual speakers. The results are interpreted in terms of the influence of L1 and L2 in third language learning in bilinguals . These results suggest that interference phenomena in L3 can be explained in terms of the acoustic nature of the sounds of L1.

1.INTRODUCTION

Accent in second or third language oral productions can be explained in terms of interference between the mother tongue and the acquired language/s. It is sometimes possible to make some predictions based on phonological descriptions but this can lead to hypotheses that do not correspond to the actual problems encountered during L2 or L3 acquisition. It then becomes necessary to characterize phonetic interference phenomena at subphonemic level, using experimental techniques. Some research along this line has been carried out by J.E. Flege and his associates but they deal with second language acquisition by monolingual speakers.

The phonetic performance of bilingual speakers has also been studied from an experimental point of view, but little is know about the pattern of interference between first, second and third language. This paper describes three experiments which aim at assessing the influence of L1 and L2 in L3 productions of bilingual speakers.

The subjects studied are either bilingual speakers having Catalan as a first language and Castilian as second language or monolingual speakers having Castilian as a first language. It has to be borne in mind that for the first group of subjects "bilingual" is a somewhat confusing designation since the subjects studied don't have the same level of proficiency in both languages, Catalan being dominant over Castilian.

2.EXPERIMENT 1

In this first experiment the production of French oral vowels by bilingual learners of French as a third language was studied .

2.1. METHOD

Ten bilingual university students of French (6 female and 4 male) read a series of Catalan, Spanish and French isolated vowels inserted in carrier sentences with a word containing the same vowel that was pronounced in isolation (e.g. " Il a dit "i" comme dans "si" "). The vowels studied were [i], [e], [ɛ], [ə], [a], [ɔ], [o], [u] for Catalan, [i], [e], [a],[o], [u] for Spanish and [i], [e], [ɛ], [a], [ɔ], [o], [u], [y], [œ], [ø], [ə] for French. Recordings were made in anechoic conditions using a Revox A77 tape recorder and a Sennheiser MD 44N1 cardioid microphone placed at constant distance from the mouth.
An acoustic analysis of a total of 240 utterances was made using a Brüel & Kjær 2033 narrow band analyzer. The frequency of the first two formants was determined from visual examination of narrow band spectra obtained using a FFT algorithm.

2.2. RESULTS

F1/F2 plottings for the male speakers are shown in figs. 1 and 2. The analysis of their French productions suggests two different problems: the series of central rounded vowels and the mid-open / mid-close pairs.

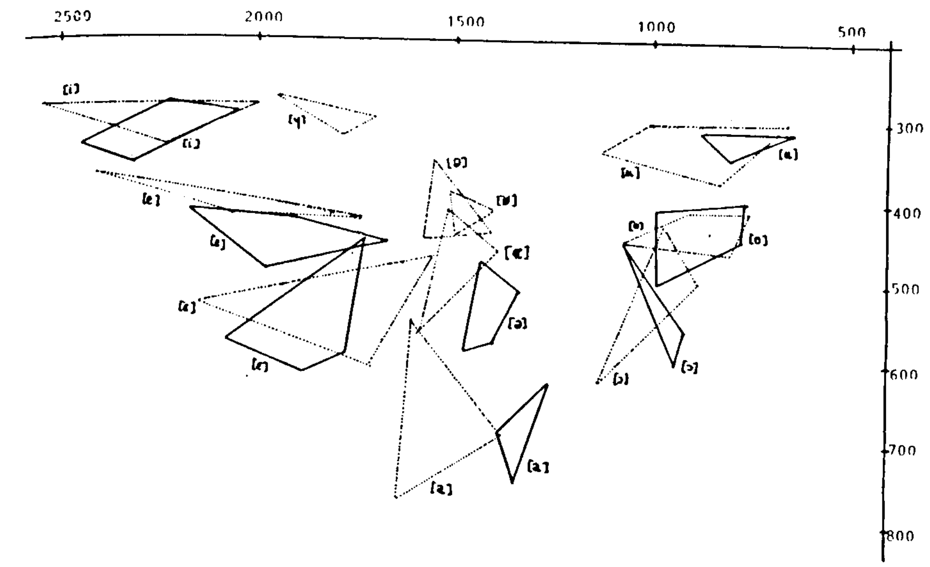


Fig 1: Catalan and French vowel productions bilingual speakers (Catalan: continuos line; French: dotted line).

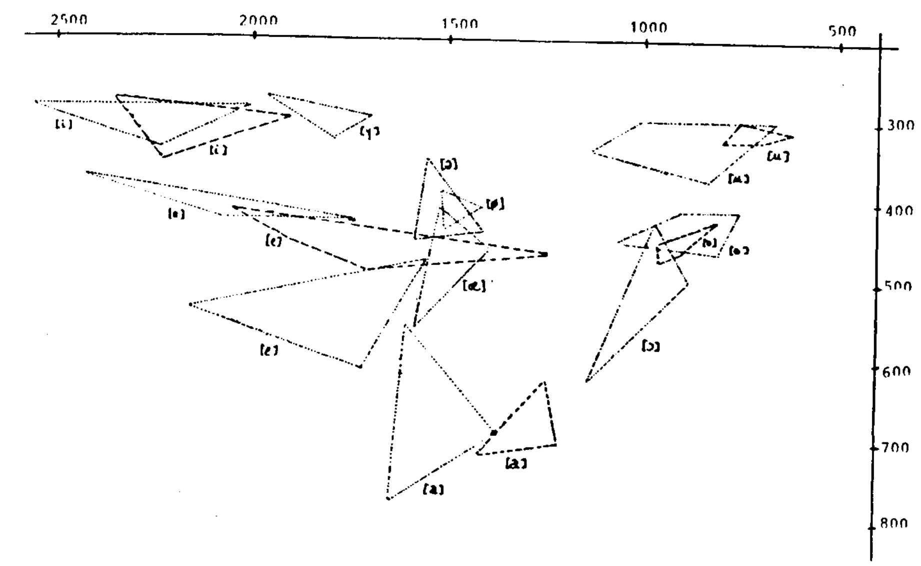


Fig. 2: Castilian and French vowel productions by bilingual speakers (Castilian: dashed line; French: dotted line)

Since central rounded vowels do not exist either in Catalan or in Spanish they tend to be clustered in a central area of the F1/F2 vowel space with no differentiation between the members of the class. In a native speaker production there is some overlapping, but always to a lesser extent than in non-native vowels. As for the [e]-[ɛ] and [o]-[ɔ] pairs, their situation in the F1/F2 plane shows the same pattern in French and in Catalan. However, the Castilian productions of bilingual speakers differ from the results obtained for native Castilian: for the bilingual speakers [e] and [o] appear in the same area as Catalan and French [e] and [o], whereas in Castilian they tend to show the same dispersion as Catalan or French [e]-[ɛ] and [o]-[ɔ] [1] [2].

3.EXPERIMENT 2

In this second experiment we tried to analyze the production of English vowels comparing the performance of bilingual and monolingual learners of English as second or third language.

3.1. METHODS

Four monolingual and five bilingual university students of English were asked to read a series of quasi-homophone words containing the vowels of Catalan, Castilian and English and embedded in carrier sentences. Recordings were made in the same conditions as in Experiment 1 and were analyzed with the same techniques. A total number of 386 utterances were measured.

3.2. RESULTS

Measurements of vowel durations on oscillograms showed that both bilingual and monolingual speakers do not make significant differences between long and short English vowels; the only pairs where differences between the mean durations were found to be significant were [i]-[i:] and [ə]-[ɜ:].

The results for vowel quality are summarized in Figs 3 and 4. It can be observed that there is a high degree of overlapping between

English central vowels, both for bilingual and monolingual speakers; Catalan speakers tend to produce the English schwa with the same acoustic characteristics as Catalan schwa.

Bilingual subjects show a better distribution of the English open/close vowels [ ]He] and [H Ho], which appear with strongly overlapped areas in the production of monolingual speakers, due to the lack of this pair in the first language.

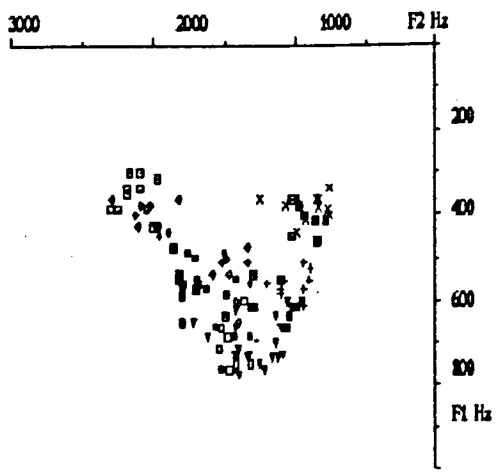


Fig. 3: English vowels by bilingual speakers

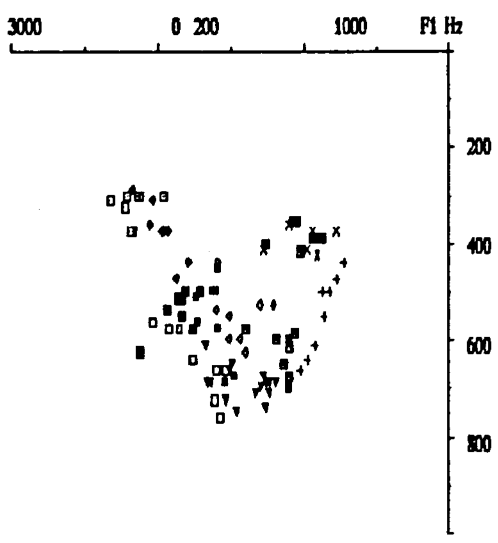


Fig. 4: English vowels by monolingual speakers

4. EXPERIMENT 3

In Experiment 3, the comparison between the performance of monolingual and bilingual speakers was extended to the fricative consonants of French. French exhibits a system of fricatives very similar to Catalan - [f], [s], [z], [ʃ], [ʒ] - while Castilian differs having [f], [θ], [s] and [x].

4.1. METHOD

Four bilingual and four monolingual speakers students of French at university level were asked to read a series of carrier sentences containing words with fricative consonants in Catalan, Castilian and French. Recordings were made under the same conditions as the previous experiments. For each fricative the following acoustic parameters were considered: frequency and intensity of upper and lower limits of acoustic energy, frequency and intensity of the two fricative formants, initial and final slope, energy spread and duration of the consonant. This gives an estimate of the spectral distribution of acoustic energy for each sound.

4.2. RESULTS

Significative differences between the three languages have only been found for the voiceless alveolar [s] [3]. This sound was found to have very similar characteristics in Catalan and Castilian native productions and in Castilian productions by bilinguals. However, both groups showed significant differences with respect to native French. French [s] has higher frequency than the Catalan or Castilian [s], and it was produced by our subjects with frequency parameter values even higher than those found for native French speakers. Thisis the same behaviour as observed by Murillo [4] and is illustrated in Figs. 4 and 5.

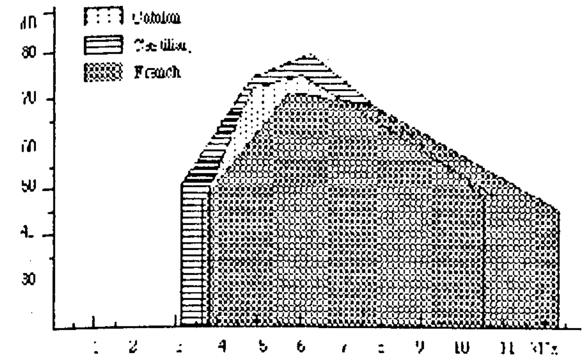


Fig. 4: Distribution of acoustic energy in Catalan, Castilian and French [s] by bilingual speakers.

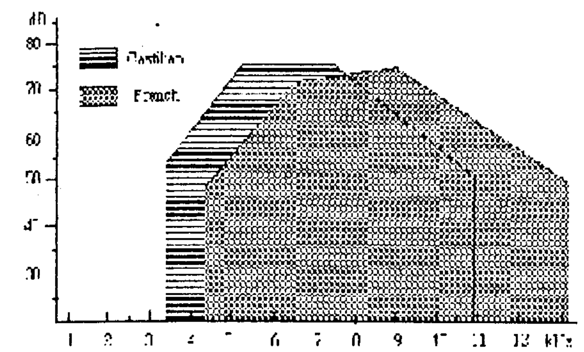


Fig. 5: Distribution of acoustic energy in Castilian and French [s] by monolingual speakers.

5. DISCUSSION

The acoustic study of foreign language vowel productions for French and English shows that bilingual speakers appear to behave in the same way when learning a third language with a complex vowel system. They tend to follow the distributional pattern of their L1 in the acoustic vowel space; the position of this same space for the L2 vowels does not seem to interfere with their L3 productions. The analysis of the L2 productions in bilingualsshows that their distribution of the vowels in the F1/F2 plane is similar in both languages, despite the fact that [e] and [o] in Castilian cover larger areas than in Catalan due to the lack of a phonologically distinctive close/open pair.

Acoustic data for the alveolar voiceless fricative [s] shows that both bilingual and monolingual speakers tend to overestimate the acoustic characteristics of the target language producing this sound with a concentration of acoustic energy at higher frequencies than those found in native speakers of French.

6. CONCLUSION

It has been experimentally shown that in the case of bilinguals learning a third language, there is no influence of their L2 in the production of L3. Interference seems to be entirely explained by the acoustic features of the sounds of their L1.

The results of acoustic analysis of L2 and L3 productions seem to suggest that, at the phonetic level,

interlanguage phenomena do not appear, since no intermediate values of the parameters measured were found.

REFERENCES

[1] Balari, S.- LListerri, J.- Poch, D. (1985) "La estructuración fonética de la materia sonora en hablantes bilingües", paper given at the XV Simposium of the *Sociedad Española de Lingüística*, Córdoba, Spain.

[2] Llisterri, J.- Poch, D. (1986) "Influence de la L1 (catalan) et de la L2 (castillan) sur l'apprentissage du système phonologique d'une troisième langue (français)", *Littérature, civilisation et objectifs de l'enseignement des langues.* Barcelona: Institut de Ciències de l'Educació de la Universitat Autònoma. pp. 153-161.

[3] Balari, S.- Llisterri, J.- Poch, D. (1987) "Structuration de la langue 3 chez les locuteurs bilingues", *Actes des Xèmes Journées Pédagogiques sur l'Enseignement du Français en Espagne.* Barcelona: Institut de Ciències de l'Educació de la Universitat Autònoma (in press).

[4] Murillo, J. (1981) *El umbral de fonologización de los sonidos agudos turbulentos del habla en español y en francés.* Unpublished PhD. Universitat Autònoma de Barcelona.

A SPEECH DISCRIMINATION TEST USING BILINGUAL COMPETING MESSAGES

J.C.WEBSTER, C.CAVE, AND M.ROSSI

CNRS Inst.de Phonetique, U. de Provence, Aix-en Provence, FRANCE

Abstract

Pairs of overlapping arithmetic statements have been recorded by native and non-native bilingual talkers using the English and French languages. In each language numerals are chosen that are easily confusable. Message pairs are counterbalanced by language, sex of talker, ear on playback, etc. such that listeners can be asked to listen to their native (non-native) language, to the (fe)male talker, left (right) ear, true (false) statement, or to the talker with(out) the foreign accent. Control tests, both messages in the same language measure the listeners' basic ability to listen selectively to one of two messages. Results on six bilingual Institut personnel confirm earlier results of a Czech/English version of the test. Only the most bilingually proficient obtain results showing both languages to be equally interferring in a selective attention task. For the others their native language prevails.

Basic Assumptions

This study starts with two basic assumptions, first, that people learn to think in a second language long before they learn to calculate in it, and that the most difficult of communicating situations is when messages compete for your attention, that is when you are trying to listen when two or more people are talking at the same time. Combining these two assumptions into a quantitative, selective attention, speech discrimination test should therefore measure a persons bilingual capability.

Test Paradigm

The basic paradigm is; use pairs of mathematical statements, either true or untrue, that overlap each other, to wit: English/English (E/E)

Fifteen and five = twenty
Fifty minus nine = six

French/French (F/F)

Cent moins dix = quatre-vingt-dix
Cinq et six = cent six

and the bilingual versions of the same sets, E/F and its equivalent F/E:

Fifteen and five = twenty . or
Cinq et six = cent six

Cent moins dix = quatre-vingt-dix
Fifty minus nine = six

Note that in all overlapping pairs, one statement is true while the other is untrue. Other parameteLcybxiM balanced in the recording and playback of the test are: one of each pair is spoken by a male voice, the other by a female voice; one is played back to the left ear, the other to the right; and half of the time the English (and the French) message is the first message. When the messages are both in the same language one is spoken by a native speaker, the other by a non- native speaker, that is with a foreign accent.

To make the statements equally difficult in the two languages numeral pairs are specifically chosen to be maximally confusable. In the above examples; in English fifteen and fifty differ by a single phoneme, they are logically and actually very confusable, as are five and nine. In French, both cinq and cent & six and dix are also maximally confusable. The first step in writing the test script is to confer with native phoneticians and map out logical and known confusions. In this manner tests have been layed out in English/Russian, English/Czech, English/Serbo-Croatian and English/German. The matrices are so chosen that any pair of languages in the sets can be matched against each other. For example, Czech/Russian, or Russian/German, or Czech/Serbo-Croatian. Although both the English/Russian and the English/Czech matrices were recorded, only the English/Czech tests were tried out on bilingual listeners, [1].

The test can be used as a simple speech discrimination test by playing back only one of the two channels and requiring a response for the "answer" to the arithmetic statement only. It was in fact tried out in this manner at schools for the deaf in Moscow and Belgrade [1].

Test Construction

The present version of the test compares English with French. Co-authors, Prof. Mario Rossi and Dr Christian Cave furnished me the list of confusable French numerals listed in Table 1.

TABLE 1

CONFUSABLE PAIRS	NUMBER OF TYPES	NUMBER OF TOKENS
1-4*,9*	2	12
1-20	1	1
2-10^,12^	2	4
3-4*	1	6
3-13^	1	2
3-30	1	1
5-7*	1	6
5-100	1	1
6-10^	1	2
11-12^	1	2
13-15^,16^	2	4
30-40^^	1	8
70-90^^	1	8
Total	16	54

KINDS OF TOKENS

*(20,30....60.80)+(1-4,9)
^(60,80)+(2-10,12)
^^(2,3....9)+(30-40)

To interpret the table note in the left column (CONFUSABLE PAIRS) that "un" is confusable with "quatre", and "neuf". In column two this is noted as two TYPES of confusions. However "un", "quatre" and "neuf" are also confusable when combined with "vingt", trente, quarante, cinquante, soixante or quatre vingt", in English twenty, thirty, forty, fifty, sixty and eighty, as shown at the bottom of the table under, KINDS OF TOKENS . Therefore these two TYPES of confusions are represented by twelve TOKENS, as noted in column 3. In line two, note that "un" is confusable with vingt, one additional TYPE and TOKEN.

Response Format

The answer sheets for these tests are in a multiple-choice format where a choice among four alternatives is required for the actual word discrimination part of the test. Table 2 shows the portion of the answer sheet that the listener would see when answering statements about the confusable pairs "un" and "quatre" and "cinq" and "cent". Choices must also be made between the operator words; plus, and, minus & less in English and between plus, et & moins in French. For the "answer" to the arithmetic statement a one-in-eight choice must be made. Scores between zero and ten points (bits) can be accumulated for correctly perceiving the arithmetic statement. Six additional points can be accumulated for the correct identity of which of the two messages was the first (second) of the overlapping pair, which was in the right (left) ear, which was spoken by the male (female), or with (without) the foreign accent and double credit for specifying whether the statment was arithmetically true or false.

Tqble 2

PROB TYPE	5 &100		35 & 45	
	1 &	4	1 &	4
5 ET 1 =	1	96	31	41
35 PLUS 4 =	4	99	34	44
45 MOINS 9 =	6	101	36	46
100 MOINS 20 =	9	104	39	49

To the left in Table 2 under "Problem Type" note the column of numerals five, thirty five, forty five and one hundred; the operator words, "et", "plus" and "moins"; and the numerals one, four, nine and twenty. Then note in the column after the equal signs the sets five and one hundred and one and four. The eight numerals in the cell below these represent all eight of the possible combinations of the larger numerals (5 & 100), the operator words and the smaller numerals (1 & 4) from the first column.

Listening Tasks

Three native French and three English speaking listeners, all with considerable knowledge and experience of both French and English served as listeners for a series of F/E, E/F, E/E and F/F tests. After extensive instructions on how to respond on the rather complex answer sheets, they were given the following series of tasks:

Task 1. On the F/E test, "respond to the message in your native language".
Task 2. On replay, "respond to the message in your non-native language".
Task 3. On the third replay, "respond to the left (right) ear message".
Task 4. On the E/E version of the test, "respond to the left (right) message".
Task 5. On the F/F version, "respond to the left (right) ear message".
Tasks 6, 7 & 8. Repeat 3, 4 and 5 answering the first (second) message.
Tasks 9, 10 & 11. Repeat 3, 4 and 5

answering the male (female) voice.
Tasks 12 & 13. Repeat 4 and 5 answering only the native (non-native) talker.
Tasks 14, 15 & 16. Repeat 3, 4 and 5 answering only that statement that is true. (No attempt was made to have them answer only the untrue statement, this would be much too difficult).

General Results

The listener's test scores can be analized in many ways. Some questions to be answered are: How well are the numerals and operator words in the mathematical statements perceived, that is, how does the test function as a word discrimination test? What type of discrimination errors are made? Do the errors tend to be random or are they concentrated on the specially chosen confusable number pairs? How accurately can the extra-acoustic and phonetic aspects of the messages be identified? Can the message content, the truth or falsity of the arithmetic statement, be be correctly ascertained? How well can the messages be selected on the basis of the acoustic, phonetic and cognitive information contained in them? Tables 3 and 4 will help to answer some of these questions.

Table 3 shows what percentage of the ten bits of information in the arithmetic statement are correctly perceived. Averaged over all listeners and listening conditions that figure is 76.18% Other aspects of the messasge that are perceived this well are the sex of the talker and the ear in which the message is heard. Whether the talker had a foreign accent or not was perceived more than half the time. Not unexpectedly the most difficult thing for the listeners, in the time allowed, (10 seconds), to record the answer was to judge whether the arithmetic statement was true or false. The most surprising shortfall was ascertaining which of the two messages was first, or second. Listeners, including the experimenters, noted that memory for the time of arrival was flighting, if the response was not recorded immediately it could not be recalled. This was not true for talker sex or accent or message localization, left vs right ear. These aspects could be answered any time before the next message arrived. The 69% overall score is ideal for tests of this type, not <u>too</u> difficult and no real problem of truncation.

Table 3
Test Scores in Percent
for various message parameters of the
Bilingual Listening Competing Messsage
Numerical Statement Test

	Listeners		
	French	English	All
Problem(10)	70.63	81.74	76.18
Acoustic(2)			
Ear	69.28	81.13	75.21
Time	38.28	63.02	50.65
Talker(2)			
Sex	69.01	83.64	76.32
Accent	51.43	69.20	60.32
Cognitive(2)	41.28	45.86	43.57
Average	63.55	75.40	69.48

Table 4 shows the average listener scores (in %) according to the task required of them. Column 1 gives the task number, the details of which are listed above. Column 2 is a short hand reference to the tasks listed above. Columns 3 and 4 list scores (in %) for the three native English-language listeners, and columns 5 and 6 list native-French-listeners scores (in %). Columns 2 and 4 are scores for statements spoken in English and columns 3 and 5 for those spoken in French.

Table 4
Test Scores in Percent
for the various listening tasks
Bilingual Listening Competing Messsage
Numerical Statement Test

		Listeners			
		English		French	
		Tests			
Tsk	Selectn Critria*	E/E F/E	F/F *F/E	E/E *F/E	F/F *F/E
1,2	N/nN*	75.00	76.69	59.38	69.79
3	Lft/Rt*	76.95	74.61	63.02	77.34
4,5	Lft/Rt	70.90	72.46	58.08	77.61
6	1/2*	75.39	78.13	53.65	66.15
7,8	1/2	73.05	69.85	62.11	77.87
10,11	Mn/Wmn	88.09	86.62	69.79	84.11
12.13	No/Act	79.45	88.28	60.94	71.62
14	Tr/Fls*	69.14	69.79	39.84	34.63
15,16	Tr/Fls	67.58	40.62	34.90	44.79
Average*		74.12	74.81	53.97	61.98
Average		75.68	72.42	57.16	71.20
ColumnAv		74.99	73.48	55.74	67.10
Grand Average			67.83		

*for the (N)ative/no(nN)ative Language task and all other tasks marked with the asterisk, the F/E test was used.

In both Tables 3 and 4 it will be noted that the "English" listeners obtain higher scores than the French listeners. This reflects the fact that as a group they were considerably more experienced in French ,than the French listeners were experienced in English. Two of the three had been married for ten years to French spouses and had resided in France the whole period The French listeners had at most spent two years in America, and one was a student who had yet to spend any time in an English speaking country. As a group the English listeners show negligible differences between scores on the English statements and scores on the French statements. By contrast the less experienced French listeners make higher scores on French statements regardless of whfether they are overlapped by other French messages or by English messages.

Concerning the types of word/numeral confusions: the majority of errors were ommission errors, but there were many cases of obvious errors among confusable pairs. These were often made to make a logically untrue statement true. The use of numerals does seem to sort out those who have really mastered the second language. Where the experieced English listeners did differ was on asnswereing true statementa as well in French as in English. Only one person could do this and he admitted he had finally learned to "calculate" only withibn the last two years, which happened to coincide with helping his young elementary school boy memorize his "arithmetic tables".

[1] J.C.Webster,"Applied Research on Competing Messages" in J. Tobias and E.D.Schubert, Eds, <u>Hearing Research and Theory</u>, Academic Press, 1983, New York

NATIVE OR ALIEN: VERIFICATION OF FOREIGN ACCENT IN THE SPEECH OF RUSSIAN LEARNERS OF ENGLISH

ANATOLY M. FEODOROV

Dept. of theoretical and experimental phonetics
Minsk State Teachers' Training College of Foreign Languages
Minsk, Byelorussia, USSR, 220034

ABSTRACT

The paper reports on an attempt to obtain some objective criteria for measurement of prosodic interference. A new technique based on rank correlation analysis is suggested to relate perceptually valid cues with acoustic features signalling foreign accent in utterances as spoken by Russian learners of English (6 males) and compared with native speakers (2 males and 2 females).
The experiment was repeated with the same speakers reading the same material after two years' period of studies.
The calculated coefficients of cross-correlation between relative vowel durations and pitch values were compared within the same speaker and with native speakers. The intraspeaker correlation proved to be much higher than interspeaker data which was in agreement with experts' rates in listening tests by a force-choice and category judgement methods.

INTRODUCTION

Most phonetic investigations on foreign accent are primarily concerned with the detection of divergencies from the phonic norms of a target language, rather than its prosodic features /3;4/. One of the reasons lies in the inadequacy and incompatibility of the existing descriptions of prosodic systems/10/.

R. Collier /2/ stresses that the linguistic description remains incomplete as long as it does not account for all perceptually relevant pitch events, but only for those that are distinctive. That is why many relevant intonation phenomena are often overlooked when they are described in terms of tone units, pitch phonemes etc.

Durational properties seem to be a sort of the resultant of a number of speech factors such as degree of prominence, phonetic makeup, syntactic structure of an utterance, pausal effects, speaker's idiosyncrasy etc. Thus rhythm is believed to be the backbone of melody, the framework on which it hangs /11/.

Speech timing control along with melody are an integral part of the speaker's linguistic competence, and the interfering effect of a first language must manifest itself as distortions of temporal and pitch structure which are perceived by native users of the language /6; 7/.

Compared to grammar learning and pronunciation of phonemes in speech flow, prosodic features of a second language are acquired later, errors persist longer and are more difficult for the learner to realize and correct, since intonation has only marginal meaning.

It is agreed that learning the phonology of a second language is a process of gradual, progressive approximation toward target language norms. This is not the case with mastering intonation, the process extremely uneven, prone to cease at early stages /4; 5; 9/. This fact suggested an idea to reproduce the experiment after two years' period of studies by employing the same subjects reading the same experimental material.

This study was designed to reveal the departures from the authentic prosody in the Russian-accented English utterances. With this object in view an attempt was made to develop a formalized technique for analysis and measurement of prosodic interference. Besides, it seemed interesting to compare the results of listening tests in which two groups of auditors - native and non-native speakers took part.

I. ACOUSTIC EXPERIMENT

Method

Stimuli and Procedure. The experimental material consisted of five sentences embedded in short dialogues and one tongue-twister. Their length varied from 6 to 20 syllables, and they were of various syntactic structure and communicative type. These sentences are given below.

(1) Could you turn your TV down a fraction?
(2) Peter Piper picked a peck of pickled peppers.
(3) A friend told me I could find some accommodation here.
(4) I'd rather have a cup of coffee if you don't mind.
(5) Yes, and it matches your scarf perfectly.

(6) I'm sorry but I seem to have mislaid your scarf.

Six Russian learners of English (all male) and four native users (two male) of the same age group naive to the purposes of the experiment were asked to solo read the test material.

The Russian learners were half-way in their five-year course of studies at the English department, the University of Leningrad. They spoke English fluently with nearly all English sounds.

The material was recorded in a sound-proof studio and then subjected to an acoustic analysis. Electronically obtained fundamental frequency trajectories were manually smoothed by continuous curves throughout the utterance. Each contour was divided into a number of regular time intervals equal for all speakers. Depending on the length of an utterance the time lag could vary anywhere from 50 to 200 ms.

Frequency measurements were taken at these points to obtain a reduced contour description that would allow point-to-point comparison between different speakers. The oscillograms of the test utterances were segmented into vocalic and consonantal segments. To facilitate this task the sentences were purposefully made up of words carrying mostly voiceless plosives and fricatives. The durations of vowels were read to an accuracy of 5 ms.

Speech rate was calculated as the ratio of overall articulation time (ms) to the number of phonemes in the ideal (careful) transcription of the utterance.

The experiment was reproduced two years later with the same learners reading the same test material under the same experimental conditions.

It is customary to assume that human perception deals with relative properties of fundamental frequency and timing by rating acoustic events within a linguistic unit. This concept conforms to rank correlation statistics to the best advantage and, in particular, Spearman rank correlation coefficient.

The coefficients were computed to analyse the degree of agreement between different productions of the same sentence by different speakers and by the same subject two years later. The obtained data were presented in correlation matrices for each utterance, pitch and durations being considered separately. Two resultant (mean) matrices for each parameter were also calculated.

In order to visualize the degree of similarity of pitch contours as well as time patterns, correlation matrices were transformed into correlation graphs through the use of an algorithm of maximum correlation.

II. RESULTS AND DISCUSSION

Table 1 summarizes the data derived from the analysis of correlation matrices for for two parameters.

Table 1. Percentage of significant correlation coefficients computed separately for pitch and timing pattern similarity between native and non-native speakers of English at two levels of confidence

	Pitch patterns		Timing patterns	
Speakers	Russian	English	Russian	English
Russian				
p = .05	48	42	83	74
p = .01	12	15	61	40
English				
p = .05		61		72
p = .01		25		56

The results presented in Table 1 clearly indicate that the subjects were able to approximate the timing organisation of the target language sufficiently well. The native speakers appear to allow greater variability of rhythmic structures than Russian speakers.

As far as pitch contours are concerned, the best agreement is observed for utterances produced by native speakers, though the group consisted of two male and two female subjects.

The difference in the percentage of significant correlations between the group of Russian learners and native speakers was found to be statistically irrelevant for pitch pattern correlation.

The data obtained from the same Russian subjects after two years of studies demonstrated that significant intraspeaker correlations accounted for 75-100 per cent of all coefficients. At the same time cross-correlation with one of the native speakers did not show any marked improvement compared to earlier performance.

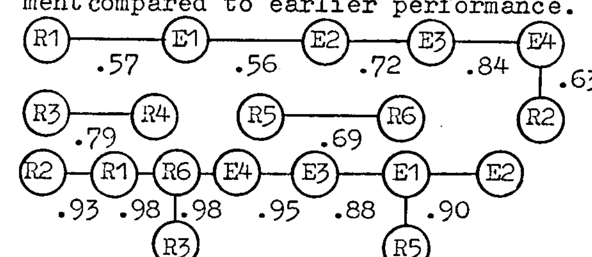


Fig.1. Correlation graphs of pitch contour (top) and timing pattern similarity: sentence 6 as spoken by English (E) and Russian subjects (R).

Examination of correlation graphs made it possible to specify some utterances as most indicative for verification of the speaker's language background. These are graphs for sentences 4, 6 andthe resultant graph, based on the mean matrix for pitch contour correlation, and graphs for sentences 3 and 6 pertaining to the temporal structure of the utterances. It is easy to see that in these graphs native speakers form clusters which suggest that cross-correlation between native-spoken utterances is greater than correlation with the other Russian-accented utterances (Fig.1).

III. AUDITORY TEST

Material, Subjects, Procedure

The same test utterances as in the previous experiment were segmented from a broader context read by the same Russian speakers and two native speakers E2 and E4 (both male). The utterances were paired with each other and ordered at random, all samples occurring in the first and second position equally. Between the first and the second member of each pair, 1-2 sec silence was inserted; each pair was repeated once, and four seconds intervened between pairs of stimuli.

The listeners were 15 Russian teachers of English phonetics and 10 British students. They were instructed to choose from each pair the sample they thought preferable as regards intonation ignoring possible occasional mistakes in sounds.

In another series of listening sessions the task of listeners was to rate the degree of similarity between two successive utterances on a five-point scale. The listeners grades were expressed in per cent for each utterance and pooled in matrices which were transformed into correlation graphs. The latter were compared with the correlation graphs of acoustic similarity obtained earlier.

IV. RESULTS AND DISCUSSION

As can be seen from Table 2, there is a good agreement between the judgments made by both groups of experts. However, the values of rank coefficient vary from 0.72 to 0.90 that suggests the same divergences from the standard intonation pattern had a different effect on two groups of listeners.

It should be noted that non-native utterances were sometimes prefered to those spoken by native speakers. These data testify to the fact that native speakers may depart sufficiently from the commonly accepted norms of their native language.

The comparison of graphs obtained for pitch contour correlation and timing similarity, on the one hand, and graphs of perceptual likeness, on the other, revealed certain isomorphism in their structure, i.e. certain clusters in one graph corresponded to analogous subgraphs in the other. By pre-assigning thresholds on graphs we were able to obtain subgraphs composed mostly of native speakers.

The subjects made no overt analysis of their reasons for prefering a stimulus but they appear to weigh up temporal and melodic factors involved in the judgment and combine them into a single response. Native listeners were found to be more responsive to the distortions of rhythmic pattern of the utterance.

Using the available graphs as the base, we selected most representative utterances with faulty rhythm and melody for comparative analysis of sentence prosody.

The comparative study of pitch models and timing patterns has enabled us to establish the following acoustic cues which contribute to the detection of Russian accent in English prosody:

(1) Russian speakers tend to level out contrast in length between phonologically long and short vowels that affects the

Table 2. Mean opinion scores (in per cent) assigned by English (E) and Russian (R) listeners in auditory tests by force-choice judgment method, and Spearman rank correlation coefficients between their responses

Speaker / Listeners	Sentences 1		2		3		4		5		6		Average percentage	
	E	R	E	R	E	R	E	R	E	R	E	R	E	R
R1	42	33	45	38	31	28	10	18	48	31	29	21	37.3	38.5
R2	71	56	50	21	50	35	49	32	55	57	47	65	51.2	56.0
R3	69	41	71	59	59	44	60	63	31	27	69	46	58.8	58.8
R4	3	26	30	33	49	15	5	27	70	50	58	31	39.6	38.5
R5	48	33	15	17	21	45	82	37	53	41	37	19	46.4	44.1
R6	72	44	85	72	62	62	78	64	39	39	54	40	65.2	67.2
E4	82	71	75	76	87	79	81	83	75	86	82	78	80.5	78.7
E2	91	98	85	80	83	91	94	80	75	71	91	91	86.7	85.1
Rank correlation coefficients, ro	0.97		0.95		0.72		0.79		0.96		0.81		0.99	

rhythmic structure of word-like phonetic units;

(2) slower overall speech rate of Russian speakers (87.91 ± 7.95 ms as against 80.3 ± 7.45 ms per sound);

(3) greater relative duration of auxiliaries and other grammatical words in an utterance;

(4) less distinct lengthening of vowels at the end of an utterance;

(5) Russian speakers are apt to lengthen excessively stressed syllables and shorten unstressed ones;

(6) timing pattern distortions result from inability of Russian learners to observe stress shifts under the influence of rhythmic tendency;

(7) pitch rise on the first pre-stressed syllable occurs much more often and is greater in magnitude;

(8) there is a strong tendency among Russian learners to use less contrastive rise-falls at the begining of an utterance;

(9) preferable use by non-native speakers of downward glides on tonic syllables;

(10) the first pitch rise occurs earlier in utterances spoken by native speakers due to a higher relative speech rate on the segments preceding the major stress.

V. CONCLUSION

As a result of the considerations presented in this paper, it appears that fluently speaking Russian learners are able to reproduce English sentence rhythm to a rather high accuracy in terms of relative durations of vowels. By contrast, sentence pitch movement proved to be much more informative for the detection of residual effects of a second language.

Intraspeaker correlations between utterances replicated by the same speaker after two years of studies were found to be greater than interspeaker correlations, especially, with native speakers. This outcome strongly suggests that in the learning process the speaker tends to adjust earlier acquired melodic prototypes to the target language by working out a prosodic idiolect.

Non-native speakers are prone to lapse into vernacular pitch and rhythm patterns unless special attention is paid to remedial exercises.

Prosodic interference seems to be caused by language-specific phonetic factors rather than phonological aspects of melody and rhythm. These phonetic peculiarities influence the authenticity of foreign language speech production to a great extent.

The present study explores only one aspect of speech performance - prepared reading aloud, and does not tackle the problem of spontaneous foreign speech.

Indications are that interlanguage prosodic interference becomes more apparent in casual speech.

REFERENCES

/1/ Adams, C. (1979): English Speech Rhythm and the Foreign Learner (Mouton, The Hague)
/2/ Collier, R.(1974): Intonation from a structural linguistic viewpoint: a criticism. Linguistics, 129, 5-28
/3/ Flege, J. (1980): Phonetic approximation in second language acquisition. Language Learning, 30, 117-134
/4/ Johansson, S.(1978): Studies of Error Gravity: Native Reactions to Errors Produced by Swedish Learners of English, 44, Acta Universitatis Gothoburgensis
/5/ Johansson, S.(1980): Another look at foreign accent and speech destortion. Revue de phonétique appliquée, Mons, 53, 35-48
/6/ Nash, R. (1972): "Phonemic and Prosodic Interference and Intelligibility", in Proc. of the 7th ICPhS, ed. A. Rigaut, (Mouton, The Hague), pp. 570-573
/7/ Phillipson, R.(1981): "Prosody Errors in English Spoken by Danes", in Nordic Prosody II, (Gleerup, Lund), pp. 271-278
/8/ Rothauser, E.et al.(1971):A comparison of preference measurement methods. JASA, 49, 1297-1308
/9/ Tahta, S., Wood, M. (1981): Foreign accents: factors relating to transfer of accent from the first language to a second language. Language and Speech, 24,265-272
/10/ Vaissière, J.(1983): "Language-Independent Prosodic Features", in Prosody: Models and Measurements, ed. A. Cutler et al.(Springer-Verlag, Berlin etc.),pp.53-66
/11/ Classe A. (1939): The Rhythm of English Prose (Blackwell, Oxford)

DIFFICULTIES IN COMPREHENSION OF L_2 INTONATION: DIAGNOSIS AND PREDICTION IN ENGLISH

MADALENA CRUZ-FERREIRA

Viken, Sweden

ABSTRACT

The paper presents a general test of intonational comprehension of English which may be used by learners of any language. It consists of sentences of English spoken with particular intonation patterns of tonality, tonicity and tone. Three alternative interpretations are provided for each sentence, the learner having to match each sentence with only one of the interpretations. It is proposed that the test may be enlarged and/or modified so as to be applicable in diagnosing and predicting difficulties with the intonation of any L_2.

INTRODUCTION

This paper presents a general test of intonational comprehension of English which, it is assumed, can be used by learners of English of any native tongue. It is an enlarged and more comprehensive version of an English and Portuguese test presented earlier to Portuguese and English speakers, respectively, in /3/. This is eminently a practical test, and theoretical discussion is reduced to a minimum. For discussion on the testing of non-native intonation see /1/, /5/, /7/, /8/ and /9/.

The reasons for the presentation of the test can be summarized in three main points:

- intonation is still the last stronghold of a foreign accent in speaking any L_2;
- intonation has only recently begun to be seriously and systematically taken into account both in the literature devoted to foreign language learning and in teaching itself;
- the test will not only help in detecting these difficulties but also, through the setting up of a typology of errors, help to correct them.

It is assumed that the first difficulty faced by non-natives with the intonation of L_2 lies in comprehension and not in production: in the first stages of learning, a faulty comprehension of intonation will determine difficulties in production, and not vice-versa. There is no point in trying to elicit a "correct" intonational form from a non-native, by imitation or otherwise, if he does not perceive it as sui-generis and specific to the foreign language. This paper presents therefore a comprehension test.

The test is proposed both as diagnostic and as predictive. Diagnostic in the sense that it enables the teacher to ascertain in a straightforward way the difficulties of the learners - what these are and where they are. For a more accurate evaluation of the results, the format of the test is such that statistical treatment of the replies of the learners is quite easy. The test is also meant to be predictive, in the sense that each sentence presented is typical of a range of other sentences where the same correspondence intonation pattern-meaning applies. One point needs explicit clarification: I do not believe that intonation works, that is, means, independently of the lexico-grammatical sentences it occurs or, rather, interacts with. I do not therefore believe that intonation can be profitably learned or taught independently of lexico-grammatical structures. This is the real sense in which I mean the word "predictive": difficulties in interpreting a sentence such as she won't drink ˇany coffee as meaning "she drinks only some types of coffee" predicts difficulties in interpreting the meaning of any sentence with the structure "not...any" and a falling-rising intonation pattern on the quantifier "any".

BACKGROUND

The intonational devices by which the different meanings are conveyed are taken to correspond to three types of intonational choice (see /6/): tonality, the division of an utterance into intonation groups; tonicity, the placement of the nucleus, or main accent, in each intonation group; and tone, the melodic shape of the nucleus, for example rising of falling.

Examples of meaning differences brought about by intonational devices in each of the three choices are:

tonality	she ˏdressed/and fed the ˋbaby
	ˏshe dressed and fed the ˋbaby
tonicity	ˏshe gave her dog ˋbiscuits
	she gave her ˋdog biscuits
tone	she won't drink ˇany coffee
	she won't drink ˋany coffee

LAYOUT OF THE TEST

A more complete version of the test is presented in /4/. The test is composed of several sentences, each with a particular intonation of English. The sentences are colloquial in style and the vocabulary used is chosen to be as simple as possible, to avoid the possibility of lexico-semantic factors affecting the interpretation of the intonation patterns.

In each sentence, the intonation pattern presented is the one which gives the sentence the least probable interpretation, if only the strict lexico-grammatical meaning of the sentence is taken into account. For example, the written sentence

she won't drink any coffee

will be given the interpretation associated with a high-falling tone rather than with the falling-rising (see /2/ for details on this).

The interpretation of the sentences as presented in the test is thus predicted to present the highest degree of difficulty for the learners. But it is also clear that this will depend to a large extent on the interactions of intonation and lexico-grammatical structure occuring in the native tongue of the learner. Knowledge of these interactions in L_1 will therefore enable the teacher to modify the test accordingly. For each test sentence, three possible interpretations are given:

a. the correct one;

b. a wrong one, but the most probable wrong answer, that is, the one corresponding most closely to the "written" form intepretation. A majority of b. replies will in principle mean that the intonation pattern of the sentence is misunderstood;

c. also a wrong interpretation but, besides, the least probable one, given mostly to ascertain the degree of randomness in the replies of the learners. In most cases c. is not in fact a possible interpretation of the sentence, and a majority of c. replies will in principle show that the intonation patterns of the sentence is not understood.

Example:

sentence	she won't drink ˇany coffee
a.	she drinks coffee, but only special types
b.	she drinks no coffee at all
c.	she prefers tea to coffee

TEST PROCEDURE

The test sentences should be presented to the learners through listening only, ideally through headphones in a language laboratory. The learners are given answer sheets where only the sentence numbers and the interpretations a., b. and c., duly randomized and re-lettered, are written.

The learners should give one answer only for each sentence and should leave no blank answers. In other words, they are forced to make a decision and one decision only as regards the meaning of each sentence. The forced choice layout is important for the establishment, after analysis of the results, of a typology of learners' errors and difficulties.

APPLICATIONS AND USES OF THE TEST

The test has been designed to typically be used as an intonational diagnosis instrument, together with other tests (grammatical, lexical, phonetic) which may be part of regular teaching instruments of evaluation. It can furthermore be used in any situation where assessment of fluency and proficiency in any language is required, and it can be supplemented by intonational production tests.

The predictive component of the test will help the teacher in establishing a typology of errors and/or difficulties, and thereby help in the setting up of adequate correction procedures.

The analysis of the typical replies of the learners will hopefully give a clear picture of the following:

what are the difficulties - that is, what typology of errors has emerged from the results;

- where the difficulties lie - that is, where do the systematic mismatches in the interaction intonation/lexico-grammatical form appear;
- why are they difficulties - that is, are the mismatches due to interference of differently meaningful pairings of intonation/lexico-grammar in L_1;
- how to counteract the difficulties - that is, what is the best way of introducing and stabilising the specific pairings of intonation/lexico-grammat in L_2, having in mind the particular L_1 of the learners.

FURTHER DEVELOPMENTS

The design of the test is such as to be easily understood and used by both learners and teachers, and its design easily lends itself to statistical treatment of the results. Its format can also be easily modified, adapted or enlarged according to the proficiency level of the learners and to the purposes of the teaching.

In the version presented in this paper, the test is meant to be used by learners of English from any mother tongue. But it is also assumed that its underlying design is suitable for use in the assessment of learners of any other language. It is hoped that the results provided by the test can be profitably used in the setting up of a typology of intonational errors and difficulties according to the L_1 of the learners, and thereby provide insights into adequate and systematic correction procedures.

REFERENCES

/1/ Anderson, K.O., Some aspects of English language interference in learning German intonation, unpublished Ph.D thesis, University of Colorado, 1970

/2/ Berkovits, R., "Are spoken surface structure ambiguities perceptually unambiguous?", Journal of Psycholinguistic Research 10: 41-56, 1981

/3/ Cruz-Ferreira, M., Non-native comprehension of intonation patterns in Portuguese and in English, unpublished Ph.D thesis, University of Manchester, 1983

/4/ Cruz-Ferreira, M., "A test for non-native comprehension of intonation in English", IRAL, to appear

/5/ Currie, K.L. & G. Yule, "A return to fundamentals in the teaching of intonation", IRAL 20 (3):228-32, 1982

/6/ Halliday, M.A.K., Intonation and grammar in British English, The Hague: Mouton, 1967

/7/ McNaught, J., The prosodic competence of a sample of French speakers of English, unpublished M.A. dissertation, University of Manchester, 1978

/8/ Pritchard, R.M.O., "The teaching of French intonation to native speakers of English", IRAL 23 (2): 117-47, 1985

/9/ Scuffil, M., Experiments in comparative intonation. A case study of English and German, Tübingen: Niemeyer, 1982

PHONEME AND TIMBRE MONITORING IN LEFT AND RIGHT CEREBROVASCULAR ACCIDENT PATIENTS

KAREN L. CHOBOR JASON W. BROWN

Department of Neurology
New York University Medical Center
The Institute for Research in Behavioral Neuroscience
New York, N. Y. (USA) 10016

Dichotic listening studies in normal subjects have indicated a right ear (left hemisphere) preference for many linguistic stimuli, including stop consonant, initial nonsense syllables (Shankweiler & Studdert-Kennedy, 1967; Studdert-Kennedy & Shankweiler, 1975), digits, and other lexical items (Kimura, 1961); as well as a left ear (right hemisphere) preference for certain nonlinguistic stimuli, including melody (Kimura, 1964), chords (Gordon, 1970), environmental sounds (Curry, 1967), and nonverbal vocalizations such as laughing and sighing (King & Kimura, 1972).

In contrast to the concept of a left hemisphere specialization for verbal material, and a right hemisphere specialization for nonverbal material, many investigators believe the left hemisphere is specialized for analytic processing and the right for holistic processing. On this view, when musical tasks share properties with speech such as temporal order, duration, simultaneity, and rhythm (Krashen, 1973), the left hemisphere is responsible for stimulus processing. Conversely, when the musical task is free of temporal constraints (i.e., not time bound), the right hemisphere is presumably processing the information in a gestalt manner. In other words, time-dependent (sequential or temporal) processing is best performed by the left hemisphere while time-independent processing is best performed by the right hemisphere (Carmon & Nachson, 1971; Albert, 1972; Gordon, 1979). This approach is consistent with an interpretation of lateral asymmetries on the basis of degree of processing (Brown 1983) rather than parallel systems or separate processing components.

There are relatively few studies on timbre and cerebral specialization, though a left ear superiority has been demonstrated with dichotic listening techniques (Gordon, 1970). Others have found a left ear superiority for limited duration only, suggesting that the ability to detect target timbres may disappear after repeated trials (Kallman & Corballis, 1975). In one study showing no significant difference between the ears for the detection of timbre (Spellacy, 1970), the intervals between dichotic presentation and recognition stimuli were 5 and 12 sec, suggesting that a different pattern of ear advantage emerges with shorter intervals, as in the case of pitch (Wyke, 1977). With shorter intervals, the procedure approaches a discrimination task, suggesting that it is the dimension of stimulus discrimination rather than the material discriminated which gives the right hemisphere effect. Thus, Mazziota, Phelps, Carson, & Kuhl, (1982) found diffuse right hemisphere PETT metabolic activation with a timbre discrimination task.

There is some evidence for selective left hemisphere involvement in phonological processing. For example, rCBF methods have demonstrated that rhyme or suffix monitoring engages left temporal regions preferentially (Maxmillian, 1982; Knopman, Rubens, Klassen, & Meyer, 1982). Conversely, Zaidel (1977) demonstrated poor phonological feature discrimination in the right but not left hemispheres of commissurotomy subjects. To date, there are no studies of phoneme monitoring in aphasics, though it has been determined that aphasics are impaired in the discrimination of phonological contrasts (Blumstein, Baker & Goodglass, 1977) and in the labeling or identifying of consonants presented in a consonant-vowel context (Basso, Casati & Vignolo, 1977). This would be of particular interest in light of evidence that phoneme monitoring involves operations that are not essential for normal language: children who have difficulty learning to read fail on such tasks, though their ability to speak and to understand spoken language is approximately normal (Liberman, 1974; Calfee, Chapman, & Vanesky, 1972). Level of reading skill, however, does not predict performance on a phonological task (Morais, 1975) though it has been suggested that performance on such tasks might identify dyslexic individuals.

METHODS

Subjects

Twenty right-handed subjects with reportedly normal hearing (as confirmed by audiological data), English as a native language, and ranging in age from 40 to 70 years were participants in this study. All subjects sustained a single, CT scan documented unilateral (10 left, 10 right) cerebral vascular accident and had no history of other neurological disorders. Left hemisphere damaged subjects included 4 nonfluent aphasics, 3 fluent aphasics and 3 total aphasics, with lesion location as follows: 4 anterior, 3 posterior and 3 anterior/posterior. Lesion location in the right hemisphere damaged subjects included 4 anterior, 5 posterior and 1 anterior/posterior. Three left hemisphere damaged and 2 right hemisphere damaged patients were deemed musically sophisticated; each had actively played a musical instrument for at least 8 years prior to CVA.

Materials and Procedure

Each subject wore a pair of Pioneer SE-305 stereo headphones and listened to recordings on an AKAI GX4000D tape recorder. Each subject was required to indicate recognition of targets by raising the hand ipsilateral to the lesion. All stimuli were prepared at the Haskins Laboratory, New Haven, Connecticut.

Language stimuli. Stimuli consisted of monosyllabic (CVC, CVCC, CCVC) words spoken by a female. Words were arranged in two 5-min blocks of 52 stimuli each, at the rate of one every 3.7 sec, with an 8-sec interval between blocks. Stimuli ranged from 850 to 1180 msec in length. Targets consisted of words beginning with the sound/b/ and foils consisted of words beginning with the sound /s/, /j/, /r/,/k/, /w/, and /m/. No target or foil ended with the sound /b/. Targets constituted 15% of the stimuli. None of the targets or foils were repeated.

Nonlanguage stimuli . Electronically generated sounds produced by an Apple computer with Syntauri software were used as stimuli. Parameter values for these stimuli are shown in Table 1. Seven different timbres were used. Each timbre was generated at four pitch levels corresponding to middle C through F above middle C. In performing the task, subjects were trained to identify one timbre as a target. Seven other timbres judged in pilot studies to be maximally different from the target served as foils. Stimulus presentations were identical to those used for phoneme monitoring. Sounds were presented in two 5-min blocks at the rate of one per 3.7 sec. Individual stimulus durations ranged from 973 to 1183 msec. Targets represented 15% of the stimuli and were presented in the same list locations as phoneme targets.

In order to motivate subjects and to ensure attention to the task, each subject was paid 15 cents for each target detected.

RESULTS

A three-factor analysis of variance with repeated measures on one factor (number of errors) was performed, with number of errors (false positives and omissions) as the dependent variable. The two between-group factors were left hemisphere damage vs. right hemisphere damage, and anterior lesion site vs. posterior lesion site. One repeated within-group factor was task stimuli (phoneme vs. timbre.

The results (Table 2) show a main effect for task stimuli ($F = 13.57$, $p = .0025$). Left hemisphere CVA patients performed poorly on the phoneme task only, and right hemisphere CVA patients exhibited the opposite effect. False positive responses for both groups of patients for both listening tasks were categorized in comparison to target stimuli. For phonemes, high acoustic frequency (/f/, /s/) and low acoustic frequency (/m/,/j/) responses were sorted; for timbres, octave and nonoctave responses were sorted. Chi square analysis revealed no pattern of false positive phoneme responses for either group of patients but a strong pattern of false positive timbre responses ($F = 1$, $p = .0065$) for left hemisphere damaged patients only, indicating that this group of patients made errors that were in octave relation to the target.

DISCUSSION

The principle finding in the study is that left brain-damaged aphasics have more difficulty with phoneme monitoring than with timbre monitoring, while the right brain-damaged nonaphasic patients show the reverse pattern. This finding appears to be material specific, since the two tasks were designed to be (1) as analogous as possible, (2) similar in such features as volume, stimulus duration and spacing, percentage of "hits," and relative distance of foils from targets; and (3) comparable in response mechanisms (ipsilateral hand).

The observation of a reciprocal performance on these analogous tasks in left and right damaged patients supports the association of phonological processing with the left hemisphere and its disruption in aphasia, and provides support for the view that phoneme monitoring involves linguistic rather than purely acoustic or attentional mechanisms. The pattern of impairment according to hemisphere damaged is also inconsistent with an interpretation of the aphasics' performance based on task complexity or degree of effort. In fact, in pilot studies with normals monitoring for two phonemes or two timbres, the latter task was judged the more difficult.

No clear relationship was found between lesion localization (anterior vs posterior) or aphasia type and performance on the phoneme monitoring task, though the number of patients was small. Severity of aphasia, as determined by Boston Diagnostic Aphasia Examination scores, was also not correlated with task performance. Furthermore, a separate analysis of patients deemed musically sophisticated prior to their strokes failed to disclose patterns deviating from the froup mean. Specifically, musically sophisticated subjects did not make errors consistent with a left hemisphere shift for timbre processing.

Of note is the fact that these tasks were employed as activation measures in a PETT study of glucose metabolism in normal subjects (Bartlett, Brown, Wolf, & Brodie, 1985). In this study, phoneme and timbre stimulation resulted in similar patterns of metabolic rates - namely, slightly greater left than right values- though regional data showed greater intersubject variability on language activation. Specifically, we did not find task-dependent metabolic asymmetries on the phoneme and timbre stimuli such as reported by Mazziota et al. (1982) for language and timbre activation. In the latter study, however, stimuli were different in material (story vs. timbre pairs), operations (listening vs. same/different judgments); and response measures (subsequent retrieval vs. motor response). When these operations are controlled, as in the present study, phoneme and timbre stimuli give similar metabolic patterns. Thus, the data indicate that lesion effects and behavioral dissociations are perhaps more sensitive than currently available metabolic correlations.

TABLE 1
Parameter Values for Timbre Stimuli

Timbre parameter	A(target)	B	C	D	E	F	G
Percussion rate	51	225	97	11	150	250	120
Percussion volume	224	222	226	222	250	250	250
Fall rate	40	40	57	45	100	80	140
Fall volume	224	0	218	0	220	80	80
Attack rate	40	225	97	189	180	250	120
Attack volume	224	225	226	227	250	250	250
Decay rate	25	17	28	19	180	40	120
Release rate	40	40	57	68	80	80	140
Release volume	0	0	0	0	0	0	0

TABLE 2
Individual Responses for Both Groups of Patients on Phoneme and Timbre Monitoring Tasks

	Correct		False positives		Omissions	
	Ph	Ti	Phoneme	Timbre	Ph	Ti
Aphasics						
A	15	15	2	0	0	0
A	3	14	22	3	12	1
A/p*	9	14	11	5	6	1
A	14	14	0	1	1	1
A/P	7	10	12	1	8	5
P	15	15	0	2	0	0
P*	15	15	0	4	0	0
P	13	12	24	3	2	3
A/P	8	10	7	2	7	5
A*	9	14	1	3	6	1
Mean	10.8	13.3	7.9	2.4	4.2	1.7
Total stimuli 104			Total targets 15			
A	13	13	0	6	2	2
A/P	13	5	0	10	2	10
P	13	15	0	2	2	0
P*	15	15	0	2	0	0
P	14	0	0	7	1	15
P	15	7	0	29	0	8
A	15	15	0	14	0	0
A	14	1	1	21	1	14
A	14	3	0	8	1	12
P*	15	15	0	0	0	0
Mean	14.1	8.9	0.1	9.9	0.9	6.1
Total stimuli 104			Total targets 15			

Note. * = Musically sophisticated; A.P., and A/P refer to anterior, posterior, or combined lesion localization.

REFERENCES

1. Albert, M.L. 1972. Auditory sequencing and left cerebral dominance for language. Neuropsychologia. 10, 245-248.
2. Bartlett, E.J. Brown, J.W., Wolf, A.P., & Brodie, J.D. 1985 Metabolic correlates of language processing in healthy right-handed male adults. Annals of Neurology (abstract), 18 (1), 119.
3. Basso, A., Casati, C., & Vignolo, L.A. 1977. Phonemic identification defects in aphasia. Cortex, 13, 84-95.
4. Blumstein, S.E., Baker, E., & Goodglass, H. 1977. Phonological factors in auditory comprehension in aphasia. Neuropsychologia, 15, 19-30.
5. Brown, J.W. 1983. Rethinking the right hemisphere. In E. Perecman (Ed.), Cognitive processing in the right hemisphere. New York: Academic Press.
6. Calfee, R., Chapman, R., & Vanesky, R. 1972. How a child needs to think to learn to read. In L. Gregg (Ed.), Cognition in learning and memory. New York: Wiley.
7. Carmon, A., & Nachson, I. 1971. Effect of unilateral brain damage on perception of temporal order. Cortex, 7, 410-418.
8. Curry, F.K.W. 1967. A comparison of left-handed and right-handed subjects on verbal and nonverbal dichotic listening tasks. Cortex, 3, 343-352.
9. Gordon, H.W. 1970. Hemispheric asymmetries in the perception of musical chords. Cortex, 6, 387-398.
10. Gordon, H.W. 1978. Left hemisphere dominance for rhythmic elements in dichotically presented melodies. Cortex, 14, 58-70.
11. Kallman, H.J. & Corballis, M.C. 1975. Ear asymmetry in reaction time to musical sounds. Perception & Psychophysics, 17, 365-370.
12. Kimura, D. 1961. Cerebral dominance and the perception of verbal stimuli. Canadian Journal of Psychology. 15, 166-171.
13. Kimura, D. 1964. Left-right differences in the perception of melodies. Quarterly Journal of Experimental Psychology. 16, 355-358.
14. King, F.D., & Kimura, D. 1972. Left-ear superiority in dichotic perception of vocal nonverbal sounds. Canadian Journal of Psychology, 26, 111-116.
15. Knopman, D.S. Rubens, A. B., Klassen, A.C. & Meyer, M.W. 1982. Regional cerebral blood flow correlates of auditory processing. Archives of Neurology, 39, 487-493.
16. Krashen, S.D. 1973. Mental abilities underlying linguistic and non-linguistic functions. Linguistics, 115, 39-55.
17. Liberman, D.Y. 1974. Experiments in syllable and phonemic segmentation in young children. Journal of Experimental Child Psychology, 18, 201-212.
18. Macmillian, X.A. 1982. Cortical blood flow asymmetries during monaural verbal stimulation. Brain and Language, 15, 1-11.
19. Mazziota, J.C. Phelps, M.E., Carson, R.E., & Kuhl, D.E. 1982. Tomographic mapping of human cerebral metabolism: Auditory stimulation. Neurology, 32, 921-937.
20. Morais, J., Cary, L., Alegria, J., & Berlelson, P. 1979. Does awareness of speech as a sequence of phones arise spontaneously? Cognition, 7, 323-331.
21. Shankweiler, D., & Studdert-Kennedy, M. 1967. Identification of consonants and vowels presented to left and right ears. Quarterly Journal of Experimental Psychology, 19, 59-63.
22. Spellacy, F. 1970. Lateral preference in the identification of patterned stimuli. Journal of the Acoustical Society of America, 67, 574-578.
23. Studdert-Kennedy, M., & Shankweiler, D. 1970 Hemispheric specialization for speech perception. Journal of the Acoustical Society of America, 48,579-594.
24. Wyke, M.A. 1977. Musical ability: A neuropsychological interpretation. In M. Critchley & R.A. Henson (Eds.), Music and the brain. Springfield, IL: Thomas.
25. Zaidel, E. 1977. Lexical organization in the right hemisphere. In P. Buser & A. Rouguel-Buser (Eds.), Cerebral correlates of conscious experience. Amsterdam: Elsevier.

ACKNOWLEDGEMENTS

Elsa Bartlett, Ed. D. assisted with aspects of the study, and Jeffrey Miller, Ph.D. provided statistical advice. Reprinted through the courtesy of Academic Press from Brain and Language, 1987.

SENTENCE INTONATION FOLLOWING UNILATERAL LEFT AND RIGHT HEMISPHERE LESION

Parth M. Bhatt

Experimental Phonetics Laboratory, Department of French,
University of Toronto, Toronto, Canada.

ABSTRACT

The results of an instrumental analysis of acoustic attributes of sentence intonation in the speech of six Francophone patients show that subjects with left anterior lesion have relatively intact range and frequency of use of F_0 movements whereas subjects with right anterior lesions show reduced F_0 range and frequency of movement.

INTRODUCTION

<u>A. Prosodic modification following left anterior lesion</u>

The question of the status of prosodic systems following unilateral brain damage has received considerably less attention than deficits of phonetic, phonemic and morphosyntactic systems. One can perhaps attribute this to the fact that the status and function of prosodic systems has remained a topic of controversy within the framework of linguistics as a whole. Prosody is often seen as a musical or emotional supplement to speech which has a minimal linguistic role. There has been a recent renewal of clinical interest in the subject of prosody following unilateral brain lesion, but the phonetic and clinical data do not as yield a clear interpretation.

There are a number of clinical observations of prosodic modification following unilateral left anterior lesion. A number of case reports describe patients whose speech output began with a period of quasi-mutism and then evolved to the utterance of a few monosyllables with modulations of intonation [1][2][3][4]. These observations suggest that in the initial stages of Broca's aphasia, prosody seem to be relatively well preserved.

However, in further stages of evolution of language disorders following left anterior lesion, specific prosodic modifications called dysprosody, appear. Dysprosody can be divided into two major categories: a) "foreign language dysprosody" (accent change) and b) "flat, discontinous speech melody".

Case reports of foreign language dysprosody are numerous. The first such case was reported by A. Pick [5]. Pick observed a young male Czech patient with a left anterior lesion, who took on a "Polish" accent because he systematically misplaced word level accentuation which, in Czech falls on the first syllable of the word, to the penultimate syllable. G. Monrad-Krohn later reported the case of a young female Norwegian patient, with a large anterior left hemisphere lesion who had recovered fluent articulation, but was unable to produce the tonal distinction which, in Norwegian differentiate "Bönder" (farmers) from "Bönner" (prayers or green beans) [6][7]. In these two cases, it is legitimate to say that the "foreign prononciation" is directly linked to a prosodic modification.

On the other hand, a number of researchers have used the term dysprosody to designate any and all cases of strange prononciations which seemed to show traces of a foreign accent, as for example, the cases reported by Alajouanine and Lhermitte [8], Cole [9], Critchley [10], Engl and Von Stockert [11][12], Nielson and McKeown [13], Pilch [14], and Whitty [15]. In most of these cases it does not seem that the accent change can be attributed to a primary and isolated modification of prosodic systems but rather to general difficulties in phonetic realization.

The second major type of prosodic disorder that has also been termed dysprosody, is associated with subjects suffering from severe non-fluent aphasia accompanied by agrammatism. H. Goodglass [16] was probably the first to use the term dysprosody to describe the flat and discontinuous quality of the melodic line in the speech of patients with severe non-fluent aphasia. For Goodglass, this type of aphasic patients are unable to use intonation to demarcate constituent boundaries.

There are relatively few instrumental studies of prosodic modification following left anterior lesion. In a corpus of spontaneous speech produced by aphasic subjects, Danly, De Villiers and Cooper [17] observed that F_0 sentence declination was present and that major falls in F_0 occured at the end of sentences. However, sentence final syllable lengthening was absent.

In a second study based on a corpus of read sentences, Danly and Shapiro [18] were able to confirm the existence of major frequency falls in sentence final position, and the absence of sentence final syllable lengthening. Furthermore, F_0 sentence declination was found to apply to smaller domains as compared to normal speech and aphasic subjects with left frontal lesions were found to make more frequency rises than normal subjects and they did not encode sentence length by choosing a high initial F_0 peak.

Ryalls [19] found that eight subjects with Broca's aphasia showed a restricted frequency range. In a second study, this author noted that subjects with large anterior left hemisphere lesions had a higher average F_0 than control subjects [20].

According to Cooper et al. [21], Broca's aphasics show a relatively flat frequency contour, their subjects did however, produce sentence final word lengthening.

In general, the portrait of Broca's aphasics shows considerable disorders in the phonetic production of fundamental frequency. F_0 is flat and discontinuous with restricted overall range. Sentence final syllable lengthening is usually absent, but sentence final intonational falls seem to be intact.

<u>B. Prosodic modification following right anterior lesion</u>

There have been relatively few clinical reports of prosodic deficits following anterior right hemisphere lesion [22]. Recently however, the syndrome of "aprosodia", which according to Ross [23] designates a selective inability to use prosody to express emotional states, and "auditory affective agnosia" [24] which designates the inability to recognize emotional information carried by voice, have lead to a renewal of research on this topic.

As far as phonetic studies are concerned, Dordain, Degos and Dordain [25] reported monotonous voice in nine of seventeen subjects with right hemisphere lesions suffering from right hemiplegia. Kent and Rosenbek [26] also report monotonous voice in two subjects with right hemisphere lesions. In an earlier study of nine subjects with right hemisphere lesion (three frontal, three parietal and three temporal) we were ourselves able to observe that the three subjects with right frontal lesion showed restricted intonational movements in phrase and sentence final position [27]. However, these subjects produced considerable lengthening in vowel duration for phrase and sentence final syllables.

INSTRUMENTAL ANALYSIS

<u>A. Subject population and speech sample</u>

The subjects for this study were six francophone, adult right-handed patients. Three patients (A, B and C) suffered from unilateral left lesions affecting the anterior portion of the left hemisphere. Three other patients (D, E and F) suffered from unilateral right lesions affecting the anterior portion of the right hemisphere. There were four female subjects, A, C, D and E, and two male subjects B and F. As for lesion etiology, subjects A, B, C and F suffered from cerebro-vascular accidents: Subject A, a thrombosis of the internal carotid artery; Subject B, an occlusion of the middle cerebral artery; Subject C, occlusion of the internal carotid artery and Subject F from an rupture of an aneurysm of the internal carotid artery. Subjects D and E both suffered from cerebral tumors, an astrocytoma and a glioma, respectively. At the time of interview Subject A was 23 years of age, Subject B, 63 years of age, Subject C, 36 years of age, Subject D, 38 years of age, Subject E, 47 years of age and Subject F, 35 years of age.

All subjects suffered from severe hemiplegia contralateral to the side of the lesion.

For subject A the clinical interview was carried out 473 days after onset of the accident, for Subject B the interview took place 637 days after onset, for Subject C, 180 days after onset, for Subject D, 17 days after onset for Subject E, 21 days after onset and for Subject F, 14 days after onset. All subjects were in stable neurological condition at the time of interview.

The speech sample submitted to instrumental analysis was drawn from the spontaneous speech section of the clinical aphasia examination battery currently in use at the Salpêtrière and St. Anne Hospitals in Paris, France. The patients were replying to questions about their illness, their profession, etc.

For each subject approximately 300 syllables of spontaneous speech were analyzed.

The speech sample for each subject was was submitted to two parallel instrumental phonetic analyses of frequency, intensity and duration. The first analysis was carried out by a digital real-time fundamental frequency analyzer and the second by a digital real-time colour spectrograph.

<u>B. Results of instrumental analysis</u>

Subjects A, B, and C, with unilateral left anterior lesions were found to have:
a) intact range of F_0 movements (Fig.1);
b) relatively frequent use of F_0 movements to indicate sentence boundaries (Fig.2);
c) relatively infrequent use of sentence final syllable lengthening (Fig.3);
d) very frequent use of pauses(Fig.4);
e) frequent use of monosyllabic or bisyllabic utterances (Fig.5).

Subjects D, E and F, suffering from unilateral right anterior lesions, when compared to subjects A, B and C, showed:
a) reduced range of F_0 (Fig.1);
b) reduced use of F_0 movements to indicate sentence boundaries (Fig.2);
c) relatively frequent use of sentence final syllable lengthening(Fig.3);
d) reduced use of pauses (Fig.4);
e) considerably less frequent use of monosyllabic and bi-syllabic accentual groups (Fig.5).

	F_0	Range	Range Coeff.
A	173Hz	150-380Hz	1.329
B	98Hz	80-190Hz	1.122
C	186Hz	135-375Hz	1.290
$\bar{x}$			1.247
D	176Hz	130-240Hz	0.625
E	149Hz	130-220Hz	0.604
F	95Hz	70-150Hz	0.842
$\bar{x}$			0.690

Range Coefficient $t=5.621$, $p(4)<0.01$

<u>Figure 1</u> Frequency range.

	Occurences of F_0 movement	
	Group Final	Sentence Final
A	23.75%	62.98%
B	40.29%	56.25%
C	26.22%	59.10%
$\bar{x}$	30.08%	59.44%
D	13.15%	20.87%
E	35.41%	25.00%
F	38.08%	38.47%
$\bar{x}$	28.88%	28.11%

Group Final F_0: $t = 0.127$, $p(4)>0.90$
Sentence Final F_0: $t = 5.535$, $p(4)<0.01$

<u>Figure 2</u>: Percentage of occurences of F_0 movement in group and sentence final position.

	Occurence of syllable lengthening	
	Group Final	Sentence Final
A	12.50%	33.31%
B	24.02%	25.00%
C	21.97%	31.81%
$\bar{x}$	19.49%	30.04%
D	47.36%	70.80%
E	60.41%	75.00%
F	64.27%	69.22%
$\bar{x}$	57.35%	71.67%

Group Final Lengthening $t = 6.081$, $p(4)<0.01$
Sentence Final Lengthening $t=13.497$, $p(4)<0.001$

<u>Figure 3</u> Percentage of occurence of syllable lengthening in group and sentence final position.

	Number of pauses	Total pause duration	Pause percentage
A	98	13631cs	60.44%
B	100	5843cs	38.59%
C	139	13675cs	64.27%
$\bar{x}$	112.33		54.43%
D	14	1422cs	19.04%
E	23	1872cs	29.81%
F	34	2576cs	39.81%
$\bar{x}$	23.66		29.55%

Number of pauses $t=6.096$, $p(4)<0.01$
Pause percentage $t=2.488$, $p(4)<0.05$

<u>Figure 4</u> Pauses.

	Mono-syllabic groups	Bi-syllabic groups	Total
A	59.34%	27.47%	86.81%
B	57.43%	20.27%	77.70%
C	53.37%	22.08%	75.45%
$\bar{x}$	56.71%	23.27%	79.99%
D	8.00%	17.70%	25.70%
E	4.69%	15.62%	20.31%
F	14.30%	12.50%	26.80%
$\bar{x}$	8.99%	15.27%	24.27%

Monosyllabic groups $t=14.360$, $p(4)<0.001$
Bisyllabic groups $t=3.033$, $p(4)<0.05$

<u>Figure 5</u> Percentage of occurence of mono- and bi-syllabic accentual groups.

DISCUSSION

While subjects with anterior left hemisphere lesions produced a high number of pauses and a large number of mono- and bi-syllabic accentual groups, they continued to use a rudimentary system of intonational marking based primarily on F_0 movement to indicate phrase and sentence boundaries.

The results also suggest that patients with unilateral left hemisphere lesion do not show a restricted F_0 range when compared to subjects with unilateral right hemisphere lesions. This observation is slightly different, but is not incompatible with previous phonetic studies which compared brain-damaged subjects to control subjects. It is however important to note that none of the patients misplaced intonational movements or produced anomalous intonational patterns. The strategy used was simple and consisted in attributing intonational rises to syllables in non-sentence final position and major falls to syllables in sentence final position.

The intonational strategies used by these two groups of subjects have important consequences for clinical analysis of language deficits following focal brain lesions and for theories of cerebral processing of speech.

In terms of the clinical analysis of Broca's aphasia with accompanying agrammatism, these results suggest that the use of the term dysprosodic, which suggests a selective prosodic deficit, to qualify the speech output of these subjects is inappropriate. Moreover, the intonational movements produced by these patients do not have simply a musical or emotional role, they serve to delimit the major constituent boundaries of the utterances.

On the other hand, subjects with right anterior lesions relied more heavily on durational attributes, as opposed to F_0 movement, to indicate the principal syntactic units of their utterances. This finding is in agreement with previous results. It is however important to note that the overall F_0 contour of these patients is not absolutely flat. The principal difference between these patients and patients A, B and C is the lack of F_0 movement in group and sentence final position where the greatest variations in F_0 usually occur. This flat final syllable, combined with a greater degree of syllable lengthening in group and sentence final position give the auditory impression of a flat voice.

As for the clinical interpretation of "aprosodia", these results suggest that the relatively flat F_0 line which is a major component of these patients verbal output may be related to a functional deficit in cerebral processing of F_0 and not to a selective disorder of emotional behaviour.

In terms of cerebral phonetic processing, the use of a rudimentary intonational strategy by subjects with left anterior lesions in the face of massive articulatory deficits, suggests that the cerebral circuits implied in the control, planning and execution of F_0 movements are functionally separate from those responsible for the planning and execution of consonant and vowel segments. Furthermore the difference in intonational behaviour of subjects according to hemispheric lateralization of lesion suggests that there is a degree of functional specialization of cerebral circuits involved in the planning and execution of F_0 movements [31]. Unilateral right anterior lesions appear to be associated with a reduction in range and frequency of occurence of F_0 movements at major constituent boundaries. Unilateral left anterior lesions appear to be associated with a severe reduction in phrase length, causing non-fluent speech output, but with intact placement and range of F_0.

REFERENCES

[1] T. ALAJOUANINE, F. LHERMITTE (1964) Non-verbal communication in aphasia, In Disorders of language, A.V.S. De Reuck and M. O'Connor (eds.), London: Churchill, 168-177.

[2] M.BOTEZ, N. CARP, L. MIHAILESCU (1968) Prosody as a means of communication in aphasia, Revue Roumaine de Neurologie, 5, 197-202.

[3] E. BRISSAUD (1901) Aphasie d'articulation sans aphasie d'intonation, Revue Neurologique, 47, 666-669.

[4] R.DE BLESER, K. POECK (1984) Aphasia with exclusively consonant-vowel recurring utterances, Advances in Neurology,42,51-7.

[5] A. PICK (1913) Die Agrammatischen Sprachstörungen, Berlin: Springer.

[6] G. MONRAD-KROHN (1947) The prosodic quality of speech and its disorders, Acta Psychiatrica et Neurologica Scandinavia, 22, 255-69.

[7] G. MONRAD-KROHN (1947) Dysprosody or altered "melody of language", Brain, 70, 405-15.

[8] T. ALAJOUANINE, F. LHERMITTE (1960) Les troubles des activités expressives du langage dans l'aphasie et leurs relations avec les apraxies, Revue Neurologique, 106, 604-633.

[9] M. COLE (1971) Dysprosody due to posterior fossa lesions, Trans. of the American Neurological Association, 96, 151-154.

[10] M. CRITCHLEY (1970) Aphasiology and other aspects of language, London: Arnold.

[11] E. ENGL, T. VON STOCKERT (1976) Ausländischer Akzent bei Aphasie, In Interdisziplinare Aspekte der Aphasieforschung, G. Peuser (ed.), Cologne: Rhineland.

[12] E. ENGL, T. VON STOCKERT (1978) Akzentverschiebungen bei Aphasie, In Brennpunkte der Patholinguistik, G. Peuser (ed.), Munich: Wilhelm Fink, 61-76.

[13] J. NIELSEN, M. McKEOWN (1961) Dysprosody: report of two cases, Bull. of the Los Angeles Neurological Society, 26, 157-8.

[14] H. PILCH (1976) Aphasische Intonationsstörungen, Saggi Neuropsicologia Infantile Psicopedagogia Riabilitazione, 2, 33-42.

[15] C. WHITTY (1964) Cortical dysarthria and dysprosody of speech, Journal of Neurology, Neurosurgery and Psychiatry, 27, 507-10.

[16] H. GOODGLASS (1968) Studies in the grammar of aphasics, In Psycholinguistics and aphasia, H. Goodglass and S. Blumstein (eds.), Baltimore: Johns Hopkins, 183-218.

[17] M. DANLY, J. DE VILLIERS, W. COOPER (1979) The control of speech prosody in Broca's aphasia, In Speech communication papers presented at the 97th annual meeting of the Acoustical Society of America, J. Wolf and D. Klatt (eds.), New York: A. S. A., 259-263.

[18] M. DANLY, B. SHAPIRO (1982) Speech prosody in Broca's aphasia, Brain and Language, 16, 171-190.

[19] J. RYALLS (1982) Intonation in Broca's aphasia, Neuropsychologia, 20, 355-360.

[20] J. RYALLS (1984) Some acoustic aspects of fundamental frequency of CVC utterances in aphasia, Phonetica, 41, 103-111.

[21] W. COOPER, C. SOARES, J. NICOL, D. MICHELOW, S. COLOSKIE (1984) Clausal intonation after unilateral brain damage, Language and Speech, 27, 17-24.

[22] M. BOTEZ, N. WERTHEIM (1959) Expressive aphasia and amusia following right frontal lesion in a right-handed man, Brain, 82, 186-202.

[23] E. ROSS (1981) The aprosodias, Archives of Neurology, 38, 561-569.

[24] D. TUCKER, R. WATSON, K. HEILMAN (1977) Discrimination and evocation of affectively intoned speech in patients with right parietal disease, Neurology, 27, 947-950.

[25] M. DORDAIN, J. DEGOS, G. DORDAIN (1971) Troubles de la voix dans les hémiplégies gauches, Revue de Laryngologie et de Rhinologie, 92, 178-188.

[26] R. KENT, J. ROSENBEK (1982) Prosodic disturbance and neurologic lesion, Brain and Language, 15, 259-291.

[27] P. BHATT (1983) Le fonctionnement du système intonatif et lésions de l'hémisphère droit, In Neuropsychologie de l'expression orale, P. Messerli, P. Lavorel and J.L. Nespoulous (eds.), Paris: Editions du C.N.R.S., 194-214.

BURST INTENSITY AS A MEANS OF ASSESSING SPEECH MOTOR PERFORMANCE IN UNINTELLIGIBLE CHILDREN

Hedwig Amorosa, Ursula von Benda, Gerd Scheimann

Max Planck Institut für Psychiatrie, Klinik
München, BRD

ABSTRACT

As part of a larger study of children with a diagnosis of specific developmental speech and/or language disorder, the burst intensity in repetitive productions of the syllable /ta/ was examined in 24 children between 4;6 and 8;0 years of age with unintelligible speech. Twenty-four children without speech/language deficits matched for age served as controls. The intraindividual variability was significantly greater in the children with unintelligible speech than in the controls. This is interpreted as an indication of a deficit in speech motor coordination. The patterns of variability in the unintelligible children differed, however, indicating an inhomogeneous group with different types of underlying motor deficits.

INTRODUCTION

Normal children can be understood by strangers at the age of 4 years. There is a group of children, however, with normal intelligence, normal hearing and otherwise normal development who are unintelligible much longer, sometimes till age 7 or 8. Recently, the phonological aspect of this disorder has been emphasized, following a trend in studies on speech and language disorders to stress cognitive and linguistic aspects and deemphasize or disregard motor functioning. In our previous studies we found that motor coordination problems contribute as much to unintelligibility in children with specific speech and language disorders as do dysgrammatism, paraphasia, and other linguistic abnormalities [1,2].

Since speech requires constant permutations and combinations of gestures in tightly defined temporal sequences, speech motor coordination can be assessed only during the act of speaking and with methods that do not interfere with this activity. We used acoustic analysis of repetitions of a simple syllable [4] to compare speech motor coordination in unintelligible children and normally developing children of the same age.

Speech can be considered as a skilled act, practiced increasingly during childhood. Two important aspects of skilled acts are speed and low variability in repetitions of the same movement. Besides other parameters, we measured the mean intensity of the release burst of the /t/ in slow and fast repetitions of the syllable /ta/. The intensity of the release burst is dependent on the intraoral pressure build-up during the stop closure and the speed of the release of the stop. Therefore, the subglottal pressure, the opening of the glottis, the closure of the velum and the seal of the stop closure all have to be controlled and related to each other to result in an overall invariant burst intensity [5,6,8]. Measurements of intraoral pressure [3] or peak airflow [7] have shown that adults have very little variation in syllable repetitions.

METHOD

Subjects

The subjects were 24 children between 4;6 and 8;0 years of age from several special schools for speech and language handicapped children. They had been selected for our study because of their unintelligible spontaneous speech. None of the children had a subnormal IQ or any hearing deficit. A detailed language assessment revealed that the children had speech and language deficits of varying types and severity, including language comprehension problems, dysgrammatism and word finding problems. The control subjects were 24 children without speech or language problems matched for age.

Procedure

The children were tested individually in a quiet room at their school. They were asked to repeat the syllable /ta/ first slowly and then as fast as possible about 20 times each. The speech was tape-recorded with a Sennheiser microphone MKE 803 and a Nagra 4.2 tape recorder. The microphone was placed in front of the child about 50 cm from the mouth. For each child the recording level was adjusted at the beginning of the recording. The tape- recorded speech was digitized at 20 kHz. The intensity level was again adjusted at the beginning of the digitization. The syllables were segmented at the release of the stop under visual and auditory control using a segmentation program developed by M. Dames on an LSI II/73.

The mean intensity of the 12.5 ms from the beginning of the stop release was calculated in dB in relation to the overall amplitude of the analog-digital converter. This procedure allows comparisons of relative mean burst intensity in individual children but not of absolute intensity between children. The difference between the burst intensity of two consecutive syllables was calculated in % of the intensity of the preceding syllable.

Statistical analysis

Since not all children produced 20 syllables, difference scores for up to 15 syllables per child were used. These

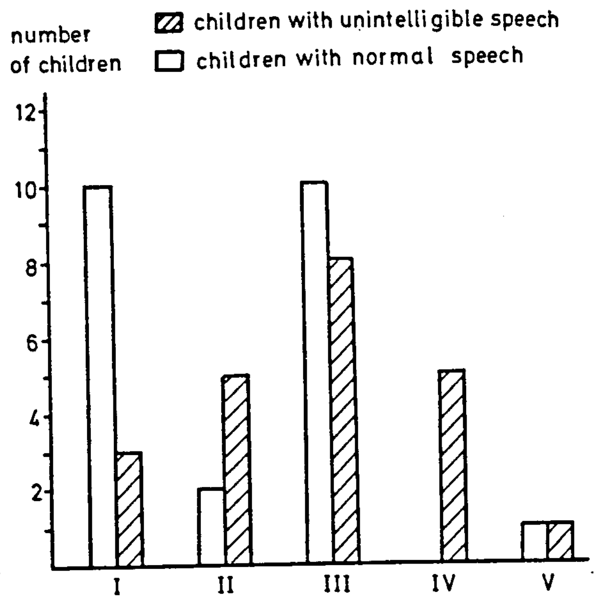


Figure 1
Classification of children into 5 clusters according to the variability of the mean intensity of the 12,5 ms after the stop release of /ta/.

scores were then grouped into 5 clusters using Ward's [9] cluster analysis. Group I included those children with less than 10% variability, Group II 10-13%, Group III 14-19%, Group IV 20-29% and Group V 30-45% variability. The Wilcoxon matched pairs test was used to test for the difference in ranks between the two groups.

RESULTS

Figure 1 shows the distribution of the children with unintelligible speech and the matched controls over the 5 clusters. In cluster I (indicating the least variability) there are 10 children from the control group and 3 children with unintelligible speech, whereas in cluster IV there are only children with speech disorders. The difference between the distributions is significant ($p<0.01$). There is no significant correlation with the variability of the mean intensity of the total syllables or the speed of fast syllable repetitions.

DISCUSSION

Zue [10] has measured the average RMS-amplitude 10-15 ms following stop release in adults. He does not give data on the variability, however. No studies on the variability of the burst intensity in children were found in the literature. If conclusions can be drawn from the low variability of intraoral air pressure [3] or peak airflow [7] for syllable repetitions, one would expect to find little variation in burst intensity in normal adults. The higher variability of the mean intensity of the first 12.5 msec of burst release in slow and fast repetition of the syllable /ta/ in children with unintelligible speech as compared to age-matched control children can be interpreted in terms of the problems these children have with motor coordination. It is most likely that they are having difficulty with the necessary and normally highly automatized constant adjustment and coordination of subglottal pressure and glottal opening, and with the nasal and oral closure. In a few of the children the high variability in the speed of the stop release might also play a role.

The group of children with unintelligible speech was not homogeneous. Some of the children were in fact able to establish control as well as the children with normal speech and language development. The high variability of the burst intensity had different explanations in different children. Some subjects were able to control the intensity for the first few syllables but seemed to lose this control later on. A few produced several syllables with very similar burst intensity and than there was a sudden change. These were children who showed involuntary choreatiform movements.

These results are compatible with our findings on phonatory abnormalities in children with unintelligible speech [1]. It seems to be the continuous adjustment and fine coordination of the different interacting systems that is the problem. Both cerebellar deficits and extrapyramidal abnormalities may be responsible for the variability in the motor output of children with unintelligible speech.

REFERENCES

[1] Amorosa, H.; von Benda, U.; Dames, M.; Schäferskűpper, P. (1986) Deficits in fine motor coordination in children with unintelligible speech. European Archives of Psychiatry and Neurological Sciences, 236, 26-30
[2] Amorosa, H.; Wagner, E. (1987) The relationship between a movement disorder of the tongue and a phonological disorder (in preparation)
[3] Brown, W.S.; McGlone, R.E. (1969) Constancy of intraoral air pressure. Folia Phoniatrica, 21, 332-339
[4] Hirose, H. (1986) Pathophysiology of motor speech disorders (dysarthria). Folia Phoniatrica, 38, 61-88
[5] Isshiki, N. (1965) Vocal intensity and air flow rate. Folia Phoniatrica, 17, 92-104
[6] Murry, T.; Schmitke, L.K. (1975) Air flow onset and variability. Folia Phoniatrica, 27, 401-409
[7] Trullinger, R.W.; Emanuel, F.W. (1983) Airflow characteristics of stop-plosive consonant productions of normal-speaking children. Journal of Speech and Hearing Research, 26, 202-208
[8] von Euler, C. (1982) Some aspects of speech breathing physiology. in: S. Grillner et.al. (eds.), Speech Motor Control. Pergamon Press, Oxford, 95-103
[9] Ward, G.H. (1963) Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236-244
[10] Zue, V.W. (1980) Acoustic characteristics of stop consonants: A controlled study. Indiana University Lingusitics Club, 1-82

INTONATION AS A POTENTIAL DIAGNOSTIC TOOL IN DEVELOPMENTAL DISORDERS OF SPEECH COMMUNICATION

Ursula von Benda   Hedwig Amorosa

Max-Planck-Institut für Psychiatrie, Klinik
Munich, Federal Republic of Germany

ABSTRACT

Speech samples obtained in 4 speech situations from 11 autistic children (7 to 17 years of age) were compared with those from speech/language disordered children and controls matched for age and IQ. The recrdings were analyzed by digital speech processing programs, the parameters assessed being f0, intensity and duration of speech segments. Analyses of variance yielded significant group differences on all three parameters, with the autistic group showing the highest intra- and interindividual variability. Discriminant analyses resulted in a clear separation of the groups. These findings support the hypothesis that intonation can be of key importance in differential diagnosis of children with developmental disorders of speech communication.

INTRODUCTION

Comparisons of morphosyntactic abilities have not resulted in statistically significant differences between verbal autistic children and speech/language disordered children [1]. However, the intonation of autistic children has consistently been described in rather impressionistic and negative terms, for example as odd, mechanical, hollow, devious or monotonous [2,3], whereas, when mentioned at all, the intonation of speech and language disordered children has been judged as normal, adequate or even "compensatory". Monotonous or idiosyncratic patterns of intonation are easily attributed to emotional disturbances because emotional aspects of verbal communication are frequently expressed solely by intonation. But intonation in its wider sense [4,5,6], acoustically a composite of the parameters f0, intensity and duration and their co-variation, serves multiple functions: on the level of the word, of the whole utterance and of the speech situation. To cite Fay and Schuler [2]: "Correct use of non-segmentals thus requires not only grammatical ability but also the ability to attend to and interpret social cues."

Normally the understanding and imitation of intonational contours precedes the acquisition of speech. Ricks [7] found some evidence that in young autistic children "patterns of babble are also impaired or abnormal". It is generally agreed that even quite intelligent older autistic children lack an intuitive understanding of intonational cues. Their literalness and lack of symbolic language might be due to this basic defect [2].

We therefore decided to compare intonational aspects of speech in autistic children, children with specific developmental speech and language disorders [8] and normally developing children. We hypothesized that measurements of f0, of intensity and of duration of speech segments would result in

1. statistically significant group differences between the autistic children on the one hand and the speech/language disordered and control children on the other;
2. Individual differences that would allow identification of each autistic child by discriminant analysis.

METHOD

Subjects

The subjects were 11 autistic children, 11 children with speech/language disorders and 11 normally developing children between 7 and 17 years of age, matched for age and IQ (Raven CPM or SPM). All of the children were of normal intelligence (±SD). They were attending schools for the language disabled or normal primary or secondary schools.

The autistic children had been diagnosed by two different psychiatrists and met Rutter's criteria for infantile autism [9].

The speech/language disordered children (SLD) met Ingram's criteria for specific speech and language disability [8].

Materials

Speech data were obtained in four different speech situations:
1. Repeating sentences (total of 14 syllables)
2. Reading sentences (total of 22 syllables)
3. Telling a story to pictures
4. Answering questions about cars.

In the latter two situations only the first 30 syllables were included in the subsequent analysis.

Procedure

Recordings were made under low noise conditions with a highly directional microphone (Sennheiser Electret Condenser Module Microphone MKE 803) placed one meter from the child's mouth. Speech signals were recorded by a NAGRA 4.2. After appropriate low pass filtering, they were digitized at a sampling rate of 20 kHz. Syllables were then segmented by visual (computer screen) and auditory feedback. F0 was determined by using a refined version of the auto-correlation-pitch-detector suggested by RABINER [10] and visually reexamined with the help of a signal editor to correct any "errors". The data were then transposed into quarter tone steps for better comparison. Intensity was measured (in dB) in relation to the individual maximum amplitude within a given speech situation.

Analyses of variance were performed to assess (a) the homogeneity of group variances (four speech situations) and (b) the homogeneity of variance of individual variances within groups (four speech situations).

Discriminant analyses were made to classify the subjects.

Variables for statistical analysis:
1. MEAN DUR/S (mean duration of syllables in msec)
2. MAX DUR/S (maximum duration of syllables in msec)
3. MIN DUR/S (minimum duration of syllables in msec)
4. MEAN F0/S (mean f0, data in quarter tones above 50 Hz)
5. MAX F0/S (maximum f0, data in quarter tones above 50 Hz)
6. MIN F0/S (minimum f0, data in quarter tones above 50 Hz)
7. MEAN INT/S (relative mean amplitude in dB)
8. MAX INT/S (relative maximum amplitude in dB)
9. MIN INT/S (relative minimum amplitude in dB)

RESULTS AND DISCUSSION

Homogeneity of group variance

For each of the 9 variables studied, the Bartlett test was used to assess the homogeneity of the estimated variance of the three groups (see Table 1).

Table 1: Bartlett test for homogeneity of group variance

Variable	CHI-SQ.	DF	Significance
1. MEAN DUR/S	19.8	2	p<.001
2. MAX DUR/S	31.3	2	p<.001
3. MIN DUR/S	4.7	2	n.s.
4. MEAN F0/S	5.9	2	p<.05
5. MAX F0/S	3.3	2	n.s.
6. MIN F0/S	2.6	2	n.s.
7. MEAN INT/S	8.2	2	p<.05
8. MAX INT/S	3.4	2	n.s.
9. MIN INT/S	0.2	2	n.s.

For variables 1 (MEAN DUR/S) and 2 (MAX DUR/S), the variances of the three groups were significantly heterogeneous (0.1% level) due to the variability in the autistic group. This was the case also for the variables 4 (MEAN F0/S) and 7 (MEAN INT/S) (at the 5% level). There was a significantly greater variability in the autistic group than in either of the other two groups. Although the difference between the control group and the SLD group was not significant for either of these variables (F-test), these two groups could be separated indirectly by comparison with the autistic group: Variable 4 (MEAN F0/S) yielded a statistically significant difference between the autistic children and the control subjects but not between the autistic children and the SLD group.

Homogeneity of variance of individual variances

We then used the Bartlett test to assess the homogeneity of variance of the individual variances within the three groups. We did this because we thought that even in those cases where homogeneous mean group variances could be assumed, homogeneity or heterogeneity of variance of individual variances might enable a clear separation of the groups.

Table 2: Bartlett test for homogeneity of variance of individual variance within the groups

Variable	CHI-SQ.	DF	Significance
1. MEAN DUR S	[illegible]	[illegible]	[illegible]
2. MAX DUR S	[illegible]	[illegible]	[illegible]
3. MIN DUR S	[illegible]	[illegible]	[illegible]
4. MEAN F0 S	[illegible]	[illegible]	n.s.
5. MAX F0 S	[illegible]	[illegible]	[illegible]
6. MIN F0 S	[illegible]	[illegible]	n.s.
7. MEAN INT S	[illegible]	[illegible]	[illegible]
8. MAX INT S	[illegible]	[illegible]	[illegible]
9. MIN INT S	[illegible]	[illegible]	[illegible]

[illegible]

[illegible]

Classification of subjects

[illegible]

and intensity but also the standard deviations of means, resulted in a very good separation of groups (see Table 3).

Table 3: Classification of subjects

Situation	Correct Classification (%) TOTAL	AUT	CON	SLD
1 Repeating sentences	94	100	91	91
2 Reading	94	100	91	90
3 Telling a story to pictures	88	91	100	64
4 Answering questions	94	83	100	91

AUT = Autistic children
CON = Control children
SLD = Speech lang. disordered children

[illegible]

Thus speech situations 1 and 2 allowed correct classification of all autistic subjects, and situations 3 and 4 of all control children. No child was classified incorrectly more than once, so that if the predominant category for a given child was used this always led to a correct assignment.

Whether this procedure will generally produce such good results must still be established by testing the model with other children meeting the same criteria.

SUMMARY

Measurements of fundamental frequency, intensity and duration of syllables in four different speech situations resulted in statistically significant differences between autistic, speech/-language disordered and normal control children (analyses of variance). Moreover, discriminant analyses allowed the assignment of each child to the correct diagnostic group.

It is noteworthy that this clear classification was possible without considering age, IQ or verbal proficiency, i.e. even with very intelligent, highly trained and/or older subjects. It appears that not only autistic children but also SLD children fail to achieve the level of proficiency that normal children do.

If analyses of other subjects meeting the same criteria yield similar results, we anticipate that in the future such evaluations of intonation with the help of digital speech processing programs may become a useful tool in differential diagnosis, even in the preverbal stage.

REFERENCES

[1] L. Swisher, M.J. Demetras, "The Expressive Language Characteristics of Autistic Children Compared with Mentally Retarded or Specific Language-Impaired Children", in: E. Schopler, G.B. Mesibov (eds.), Communication Problems in Autism, Plenum Press, New York, 1985

[2] D. Fay, A.L. Schuler, "Emerging Language in Autistic Children", in: R.R. Schiefelbusch (ed.), Language Intervention Series, Vol. 5, E. Arnold, London, 1980:43 f.

[3] C.A.M. Baltaxe, J.Q. Simmons, "Prosodic Development in Normal and Autistic Children" in: E. Schopler, G.B. Mesibov (eds.), Communication Problems in Autism, Plenum Press, New York, 1985

[4] D. Crystal, "The Intonation System of English", in: D.L. Bolinger (ed.), Intonation, Penguin Books, Baltimore, 1972

[5] V.A. Artemov, "Intonation and Prosody", Phonetica, 35/6, 301-339, 1978

[6] E. Stock, "Untersuchungen zu Form, Bedeutung und Funktion der Intonation im Deutschen", Schriften zur Phonetik, Sprachwissenschaft und Kommunikationsforschung, No. 18, Akademie Verlag, Berlin, 1980

[7] D.M. Ricks, "Vocal Communication in Pre-verbal, Normal and Autistic Children", in: N. O'Connor (ed.), Language, Cognitive deficits and Retardation, Butterworths, London, 1975

[8] T.T. Ingram, "The Classification of Speech and Language Disorders in Young Children", in: M. Rutter, J.A.M. Martin (eds.), The Child with Delayed Speech, Clinics in Developmental Medicine, No. 43, Heinemann, London, 1972

[9] M. Rutter, E. Schopler, "Autism: A Reappraisal of Concepts and Treatment", Plenum Press, New York, 1978

[10] L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-25, No. 1 24-33, 1977

[11] P. Beutel, H. Küffner, W. Schubö, "SPSS8" (Statistik-Programm-System für die Sozialwissenschaften), Fischer Verlag, Stuttgart, 1980

SPEECH PATHOLOGY IN INFANTS SUFFERING FROM INFANTILE CEREBRAL PALSY

Y.V.ROMANENKO

Laboratory of Speech Pathology
Serbsky All-Union Research Institute for General and Forensic Psychiatry
Moscow, USSR, 119839

ABSTRACT

We observed 110 infants suffering from cerebral paralysis. The aim of the present investigation is to study infant speech formation under cerebral pathology. Clinical and psychopedagogical methods were used in the investigation.

The early period of the infant development is of great importance for the normal speech formation. It is conditioned by the peculiarities of the infant brain development, optimal periods of the maturing of the speech function system as well as by its abilities to compensate disturbed functions. There is a considerable number of works devoted to the problem of speech formation in normal development, whereas the question of the development of this system under pathology has not been studied enough so far. This trend is presented in the works of E.M. Mastyuchova /1/, E.F. Archipova /2/, M. Cass /3/.

We observed 110 infants suffering from cerebral palsy. The aim of the present investigation is to study infant speech formation under cerebral pathology.

In the investigation we used clinical and psychopedagogical methods.

Infant cerebral palsy is a polyaetiological illness of the central nervous system which appears in the pre- and natal period of the infant's development and is characterized by the affection of motor and psychic spheres. According to the data of different authors speech affections are found in 70-80 % of cases of infants suffering from this illness.

The first year of the infant's life is conventionally called a pre-speech period the initial stage of which is the cry. Infants with cerebral pathology may have no cry or have a weak constrained cry which is connected with the pathology of the intrauterine period or asphyxia during birth. Normally appearing in the first weeks of life sounds appear with a considerable delay under cerebral pathology. In this case they are rare and monotonous. The early stage of baby-talk which under normal conditions appear at the age of 2.5 months develops spontaneously only at the age of 4-5 months and sometimes of one year of the infant's life under pathology. Besides the temporal delay of this stage of baby-talk development insufficient sounds melodiousness, rudimentary character of sounds realization and unmodularity take place. The main composition of the early stage of baby-talk is formed by the consonant sounds of indistinct locality - approximate vowels "а", "э", "ы". Infants for a long time stay at the period of articulatory movements realization which takes its course independently of the infant's hearing. They pass over to the next stage (autoecholalia and echolalia) with great difficulty and delay.

Baby-talk is usually delayed and starts at 9-11 months and sometimes even later - at the age of 1.5 year. The baby-talk is often poor in the sound composition. Most frequent are bylabial ("п", "б") and backlingual ("к", "г") sounds, less frequent are alveolar ("т", "д") sounds. Even during favourable development the baby-talk stage is characterized by fragmentariness, poorness of sound complexes and little activity. The stage may last more than one year.

Conventionally the end of the first year of the infant's life is marked by the speech period formation. Children with cerebral paralysis have it at the age from 1 year and 2-3 months to 2-3 years which depends not only upon the level of the psychic development of the child but also on the severity of the speech-motor pathology. The retraced dependence of the speech transition period on the moment of the baby-talk points out to its great prognosis value.

Infant's speech formation under cerebral paralysis generally follows the main principles of speech formation in the norm, although it has its own peculiarities. They are - longer periods of acquiring separate groups of sounds and dependence of speech formation on severity and affected locality of the articulatory apparatus. Infant speech-motor images of the articulatory apparatus under cerebral pathology do not serve as a necessary basis for the auditory perception of sounds, as it is observed in the norm. Hearing under these conditions hinders, instead of stimulating, speech formation.

Substitutes acquire the same pathological character - in the norm they help in the transmission from one sound to another, whereas in this case they also play a hampering role. Substitutes are not of constant character, they often change.

The sound composition of infant speech is often characterized by the presence of one or two groups of sounds (according to the manner and place of articulation) which depends on the locality and character of affection of the articulatory organs. Thus if the affection of the tongue muscles prevails labial sounds are mainly present, while the lip muscles affection gives an opportunity for the formation of lingual sounds. As a rule, we come across a mixed type of pathology which affects all the muscles of the articulatory organs, although some areas are characterized by more explicit pathological changes as compared to the others.

The affections described above at the early age are characterized as a delay in prespeech or speech development which may eventually transform into a speech breach and make the speech communication of the child impossible.

/1/ - E. Mastyuchova "Clinical picture and rehabilitation therapy of cerebral palsy in infancy", Medicine, 1972.
/2/ - E. Archipova "The prespeech period peculiarities in infant cerebral palsy", MSPI, 1979.
/3/ - M. Cass "Speech habituation in cerebral palsy", Harver Publishing Company, 1965.

NEW LARYNGOGRAMS OF THE SINGING VOICE

DAVID M HOWARD

GEOFFREY A LINDSEY

Department of Phonetics and Linguistics
University College London
Great Britain

ABSTRACT

Laryngographic techniques evolved in speech analysis are extended in the present work to the analysis of the singing voice. Attention is focused on laryngographically derived measures of vocal fold open-phase times. Specifically, the measure of open quotient (open phase time over whole period time) provides a quantitative parameter for the characterisation of voice production differences between speakers, trained singers and untrained singers.

INTRODUCTION

The laryngograph [1] has been used for many years as a tool for the analysis of normal and pathological speech as well as of the singing voice. More recently, detailed studies have investigated changes in the laryngograph output waveform (Lx) on a cycle-by-cycle basis, with a view to correlating these with the acoustic output from the vocal tract [2]. The present work is designed to develop a new series of laryngographically based plots which can be utilised on a routine basis for speaking and singing voice research and, potentially, in visual displays to be developed to give feedback in singing/speaking voice production training.

DATA, SUBJECTS, AND RESULTS

Four adult male singers took part in the experiment. Two are 'trained', having had formal voice training and extensive solo performance experience; one of these is a baritone (GW) and the other a tenor (SB). The other two have choral singing experience and vocal ranges in the middle (baritonal) range for men; one of these (DH) is an experienced amateur musician who has received vocal training for a short period, and the other (GL) is untrained. GL made recordings in both a 'natural,' informal style and in a quasi-trained style which he adopts for choral performance.

The subjects were digitally (PCM) recorded onto videotape in the anechoic room at UCL, with the output from a high quality condenser microphone (Sp) on one channel and the laryngograph output waveform (Lx) on the other.

The data consisted of:
1) a reading of a phonetically balanced passage lasting approximately two minutes;
2) five monophthongal vowels, in the environments /b_d/ and /m_n/, spoken with falling intonation and then sung on C (256Hz), C (128Hz), E (330Hz) and E (165Hz);
3) major scales exhibiting each singer's range, sung on the vowel /a/; and
4) a performance of "God Save the Queen" starting on G (192Hz).

The analyses consist of the following (see Figs. 1-6):
a) the speech pressure waveform (Sp);
b) the laryngograph output waveform (Lx), derived by measuring the current passing through the throat between two voltage driven electrodes placed on the wings of the thyroid cartilage; the peaks thus correspond to the maxima of vocal fold closure and the valleys to maxima of glottal opening, in each cycle [1];
c) vertical period markers derived from Lx (Tx) [3];
d) a plot of the logarithm of fundamental frequency (derived on a period by period basis from Tx) against time (Fx); and
e) two plots of the open quotient (OQ), which is defined as the duration of the open phase of each cycle divided by the duration of the whole of that cycle, calculated from Lx by different methods -- the first method (OQ1) takes the point of vocal fold closure in a glottal cycle to be the peak of the differentiated Lx waveform, and the point of opening to be the minimum in the differentiated Lx waveform; whilst the second (OQ2) takes the upper 70% of the peak-to-peak amplitude of each cycle to represent closure of the glottis and the lower 30% to represent an open glottis (see [3] for a full description)

Figs. 2 to 6 show these analyses plotted for the first note of the final occurrence of the word "God" (E 330Hz) in the sung performances of "God Save the Queen" by each of the subjects. Fig. 1 shows equivalent plots for the vowel [a] spoken by a female subject with high fall intonation. In each figure pertaining to singing, the OQ plots derived by both methods remain relatively steady; mean values for the portion shown are tabulated in Fig. 7. This table also gives summary statistics relating to the speaking voice of each subject based on the reading of the passage (section 1 of the recording) which have been derived from a second order fundamental frequency distribution plot [4]. This plot consists of a histogram of the number of consecutive pairs of Tx period values which fall within the same histogram 'bin'. The summarised statistics show the number of such pairs in the plot for that subject (under the heading "samples"), the Fx range at the 0.1% probability level, and the modal Fx value. Mean values of OQ calcutated by the two methods are also given in the table.

DISCUSSION AND CONCLUSIONS

In spoken data, evidence has been found [2] for a 'preferred' value of Fx towards the lower end of a speaker's overall speaking range, at which the vocal folds vibrate with optimal efficiency and vigour. At this value of Fx, open and closed phases are approximately equal and the peak-to-peak amplitude of the Lx waveform tends to be at its maximum. Independent evidence that this is indeed a preferred frequency of vocal fold vibration is found in Fx distributions for stretches of continuous speech which have their modes near this Fx value. It is plausible to suggest that this Fx value is utilised by the speaker as a departure/arrival point for prosodic excursions and as a locus for neutral, non-pitch prominent syllables. Fig. 1 illustrates this for the speech of a woman (EA), whose Fx range is measured as 118Hz to 371Hz and whose Fx mode is 147Hz (Fig. 7); this speaker's OQ1 and OQ2 values pass through the 50% level around the point at which her Fx contour passes through 147Hz.

It has also been found [3] that for many speakers closed phase duration varies considerably more with Fx than does open phase duration. That is, the open phase remains comparatively constant while the closed phase is shortened as Fx rises and lengthened as Fx falls. Thus the closed phase is roughly the inverse of Fx, while the open phase is rather steadier. This can be seen in the plots for our speaker (Fig. 1) where OQ values, calculated by both methods, tend to fall with the falling Fx.

The majority of a singer's pitch range makes use of Fx values which are high relative to the normal speaking pitch range. In view of the preceding observations on speech, one might expect raised OQ values at sung pitches at the higher end of the range. This does not appear to be the case with our trained singers. Figs. 2-6 show OQ values for the note E (330Hz), which is well above the upper limit of the measured Fx speaking range for each subject: Fig. 7 gives a summary of their speaking ranges and modes. The singers are ordered by experience, and our speaker (EA) is included as the final table entry for comparison. This ordering of subjects corresponds closely to the trends in OQ1 and OQ2 values for the singers.

Subjects GW and SB are professionally trained with many years solo performance experience. GW is also a singing teacher. In both cases the mean OQ values, calculated by both methods, are markedly lower (30% to 38%) than the values found for the other subjects (49% to 75%). Subject DH, who has had some singing training, and extensive choral experience at an amateur level has mean OQ values close to 52%. Subject GL, who has had no formal singing training, recorded productions in two maners, an informal style (GL(U)) and a quasi-trained style (GL(T)), and it is clear that he attains more appropriate OQ values when he adopts his choral style. The mean OQ values given for our speaker (EA) must be interpreted wih reference to Fig. 1, since it is clear that the OQ values vary over a wide range with Fx change: her OQ values are above 50% during the first portion of the utterance and they descend below 50% towards the end as her Fx is lowered.

These data suggest that there is a clear trend towards lowered OQ values with increased singing training and performance experience. This is of note because speech data suggest that raised

OQ values correlate with raised Fx, and therefore lowered OQ in singing must presumably be a direct effect of a trained style of singing. This implies that an important mechanism involved in note productions by a trained singer is the considerable lengthening of the closed phase in each larynx cycle with respect to the open phase. This has two main consequences: firstly the voice quality becomes less breathy; and secondly the longer closed phase ensures more prolonged substantial acoustic output from the vocal tract as the coupling-in of the subglottal cavities (and the associated increase in acoustic damping) occupies less of the cycle. Thus the singer makes use of a natural acoustic consequence of an action which presumably requires no additional pulmonic energy, to achieve this increase in output.

Future work in this area will include a wider range of subjects with various levels of singing training and experience in order to evaluate the robustness of these measures. The possibility also exists of a new form of visual display to aid the singer, which could complement the singing assessment and development (SINGAD) system currently aimed at note pitching by children from five years upwards [5]. This system makes use of a microcomputer and a specially developed hardware interface, based on [6], which estimates fundamental frequency from an acoustic input; it also allows work on vibrato, note onset and offset, and pitch systems of different musical traditions. A new OQ component along the lines discussed above, would make a significant contribution towards establishing a comprehensive and coherent tool for students of singing.

ACKNOWLEDGEMENTS

The authors would like to thank Graham Welch and Simon Bainbridge for their freely given time in making recordings, Peter Davies for his OQ analysis package and his help with software modifications, and David Smith for the Tx processing system.

REFERENCES

[1] Fourcin, A.J., and Abberton, E.R.M. (1971). "First applications of a new laryngograph", Medical and Biological Review, 21, 172-182.

[2] Lindsey, G., Davies, P., and Fourcin, A. (1986). "Laryngeal coarticulation effects in English VCV sequences", IEE Conference Proceedings 258, 99-103.

[3] Davies, P., Lindsey, G., Fuller, H., and Fourcin, A.J. (1986). "Variation in glottal open and closed phase for speakers of English", Proceedings of the Institute of Acoustics, 8, 539-546.

[4] Fourcin, A.J. (1981). "Laryngographic assessment of phonatory function", The American Speech Language Hearing Association (ASHA) Reports, 11, 116-127.

[5] Howard, D.M., Welch, G.F., Gibbon, R.R., and Bootle, C.M. (1987). "The assessment and development of singing ability - initial results with a new system", Proceedings of the Institute of Acoustics, 9, in press.

[6] Howard, D.M. and Fourcin, A.J. (1983). "Instantaneous voice period estimation for cochlear stimulation", Electronics Letters, 19, 76-78.

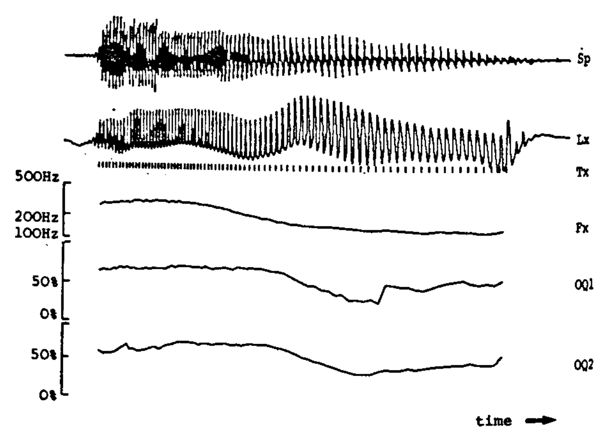


FIGURE 1: Subject EA spoken utterance [a] on a low fall

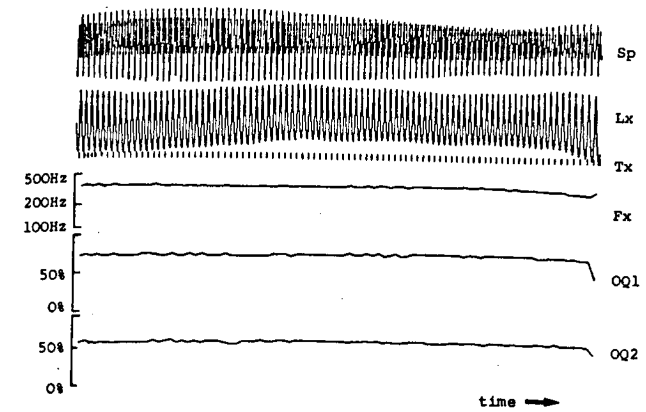


FIGURE 2: Subject GL(U) sung note: (GL in 'natural' style)

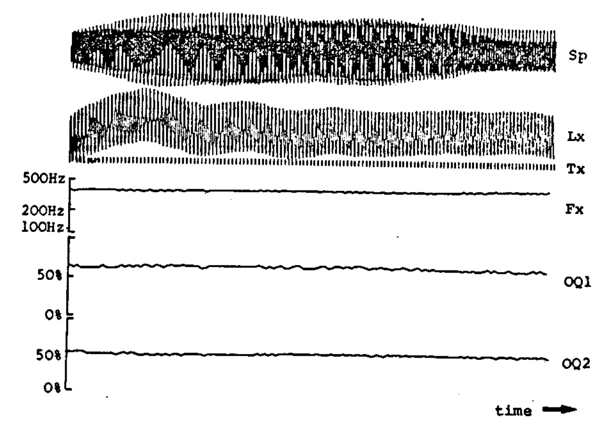


FIGURE 3: Subject GL(T) sung note: (GL in quasi-trained style)

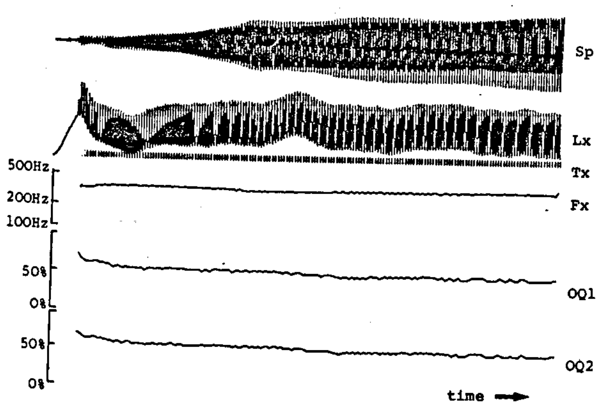


FIGURE 4: Subject DH sung note: (slightly trained baritone)

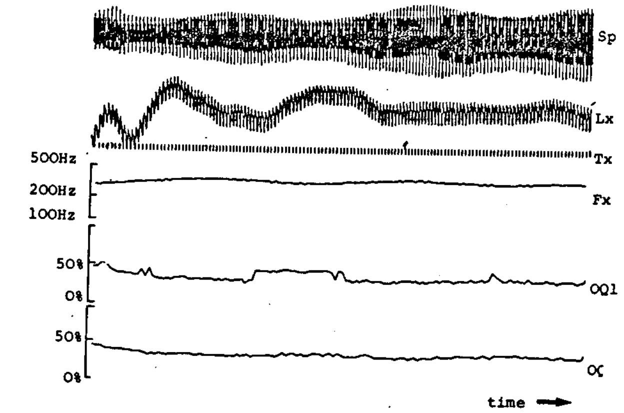


FIGURE 5: Subject SB sung note: (trained tenor)

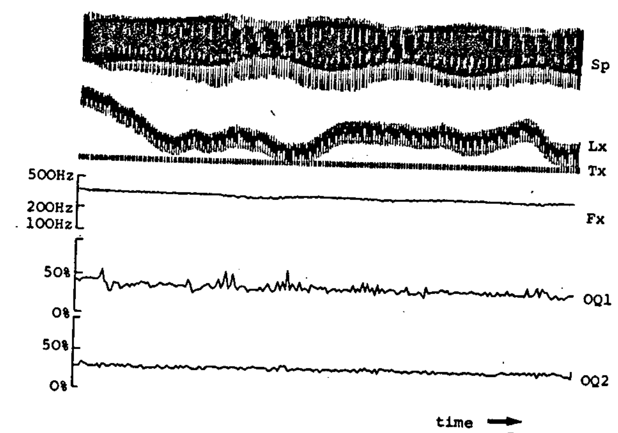


FIGURE 6: Subject GW sung note: (trained baritone)

	2nd order Fx distribution statistics - read passage			Open Quotient mean values	
Subject	samples	0.1% range	Mode	OQ1	OQ2
GW	2529	89Hz-312Hz	149Hz	37.5%	38.1%
SB	2539	104Hz-200Hz	123Hz	36.9%	35.5%
DH	2694	104Hz-288Hz	162Hz	52.1%	52.5%
GL(T)	3181	82Hz-312Hz	116Hz	64.6%	49.8%
GL(U)	3181	82Hz-312Hz	116Hz	74.6%	60.3%
EA	2758	118Hz-371Hz	147Hz	60.8%	59.4%

NOTES: OQ1 measured by differentiated Lx method
OQ2 measured by 70% to 30% Lx amplitude method
OQ measures for utterances in figures 2-6

FIGURE 7: Table of summary speech Fx statistics and mean open quotient values for all subjects

LES MOUVEMENTS DU VOILE DU PALAIS DANS LE CHANT

Nicole Scotto di Carlo & Denis Autesserre

Institut de Phonétique - U.A. 261. CNRS.

Université de Provence. France.

RESUME

Dans l'enseignement du chant, une tradition datant de plusieurs siècles, insiste sur l'importance du soulèvement du voile du palais pour l'émission de l'aigu. Or, tous nos examens physiologiques ont révélé que dans ce registre il était relativement abaissé.
Nous tentons d'expliquer cette contradiction et de comprendre comment une telle erreur dans l'appréciation d'un mouvement a pu être commise par des sujets dont les sensibilités internes sont extrêmement développées.

I. - INTRODUCTION

Depuis le XVIIe siècle où sont parus les premiers ouvrages sérieux de technique vocale, jusqu'à nos jours, une place prépondérante a toujours été accordée au voile du palais et au rôle qu'il joue dans le chant, en particulier pour l'émission de l'aigu.
Pendant trois siècles, on a écrit et enseigné que le voile du palais devait se soulever progressivement du grave à l'aigu, mais à notre connaissance, aucune étude systématique des mouvements du voile du palais dans le chant n'a encore été réalisée pour vérifier cette affirmation. Il nous a donc paru utile de combler cette lacune.

II. - PROTOCOLE EXPERIMENTAL

A. - <u>Enregistrements</u> :

Six chanteurs professionnels représentant les principales catégories vocales ont été soumis à des examens xéroradiographiques et endoscopiques avec synchronisation sonore.

1. Les clichés radiologiques ont été pris lors de l'émission de quatre voyelles tenues : /i/, /a/, /u/, /ã/, en voix parlée puis en voix chantée, dans le grave, le médium et l'aigu.
2. L'examen endoscopique a été enregistré en vidéo à l'aide d'un fibroscope souple introduit par voie nasale entre le cornet inférieur et le cornet moyen de manière à surplomber le cavum au niveau de l'apophyse ptérygoïde.

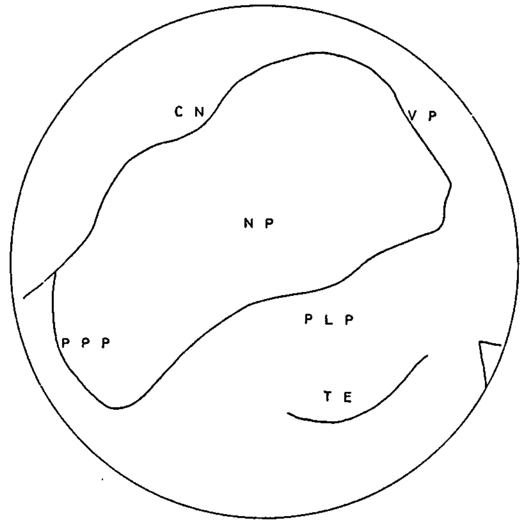
Les sujets ont réalisé les quinze voyelles du français en voix parlée, puis en voix chantée dans les trois principaux registres et sur les mêmes notes que celles sélectionnées pour l'examen radiologique.
Des calques de chaque trame (50 par seconde) ont été ensuite dessinés.

B. - <u>Paramètres analysés</u> :

1. Sur les calques radiologiques,
   - le degré de relèvement du voile a été estimé à partir de la ligne bi-spinale prolongée jusqu'à l'arc antérieur de l'atlas.
   - la fermeture vélo-pharyngée est mesurée dans la région de constriction maximale située entre le dos du voile et la paroi postérieure du pharynx (Fig. 1 B).
2. Les données endoscopiques précisent les renseignements fournis par les clichés radiologiques, sur les mouvements antéro-postérieurs du voile et apportent une information complémentaire en ce qui concerne les mouvements des parois pharyngiennes. Toutefois, les mesures des documents endoscopiques ne peuvent avoir qu'une valeur relative en raison des déformations optiques (de type grand angulaire) et des déplacements éventuels du fibroscope en cours d'examen.
   Sur les calques endoscopiques,
   - les mouvements antéro-postérieurs du voile sont mesurés par référence à un axe para-médian qui relie le centre du voile à la paroi postérieure du pharynx;
   - les déplacements des parois latérales du pharynx ont été mesurées par référence à une perpendiculaire à l'axe para-médian au niveau de la constriction maximale (Fig. 1 A).

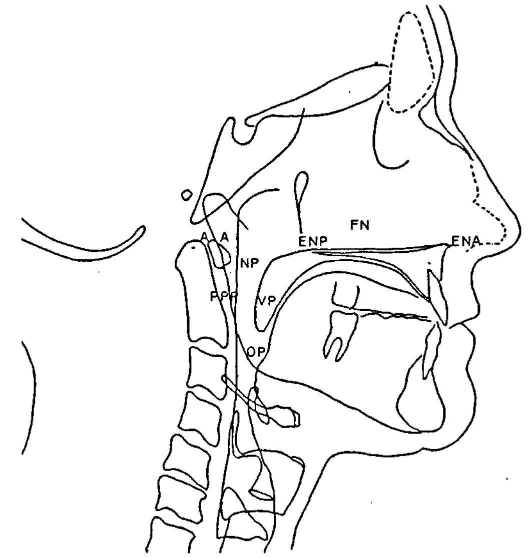
III. - RESULTATS

Dans le cadre limité de cette étude, nous n'avons retenu que deux voyelles extrêmes en ce qui concerne la position du voile :
/i/ voyelle orale, fermée, tendue, qui correspond dans la parole à un soulèvement maximum du voile et
/ã/ voyelle nasale, ouverte, détendue, qui entraîne son abaissement.
Ces voyelles ont été émises dans les deux registres extrêmes (sur si2 dans le grave et sur si4 dans l'aigu) par un soprano professionnel. D'autre part, afin de disposer d'un élément de référence en ce qui concerne l'abaissement maximal du voile, nous avons également étudié des images radiologiques et



N P : NASO-PHARYNX
PLP : PAROI LATERALE DU PHARYNX
PPP : PAROI POSTERIEURE DU PHARYNX
V P : VOILE DU PALAIS
C N : CLOISON NASALE
T E : TROMPE D'EUSTACHE

FIGURE 1A



A A : ARC ANTERIEUR DE L'ATLAS
ENA : EPINE NASALE ANTERIEURE
ENP : EPINE NASALE POSTERIEURE
F N : FOSSES NASALES
N P : NASO-PHARYNX
O P : ORO-PHARYNX

FIGURE 1B

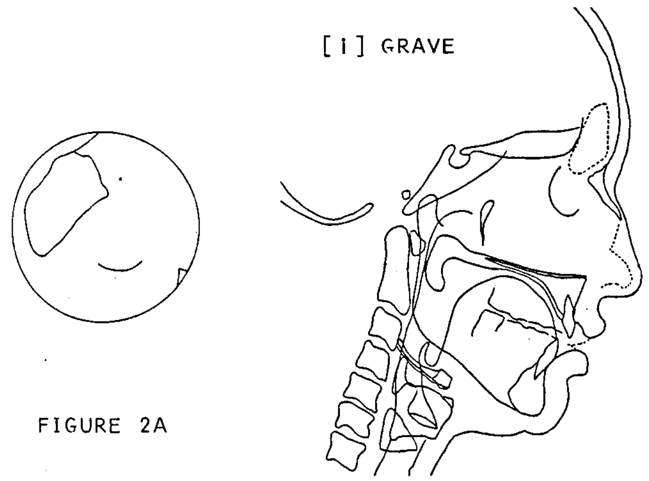


FIGURE 2A

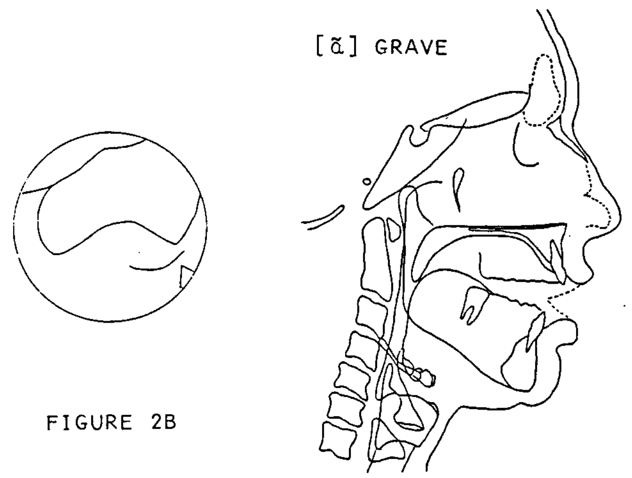


FIGURE 2B

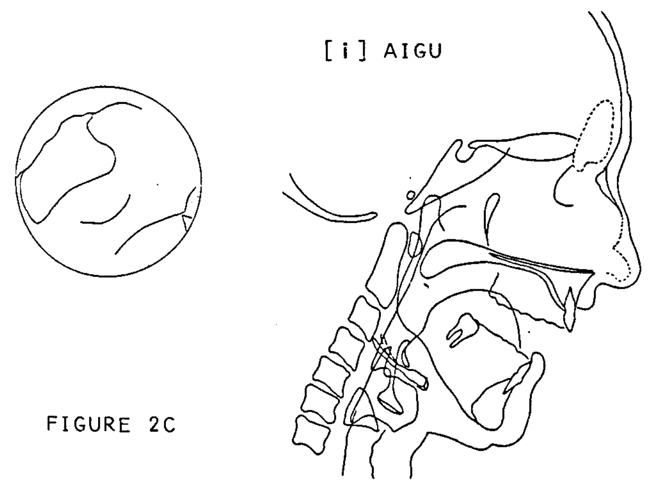


FIGURE 2C

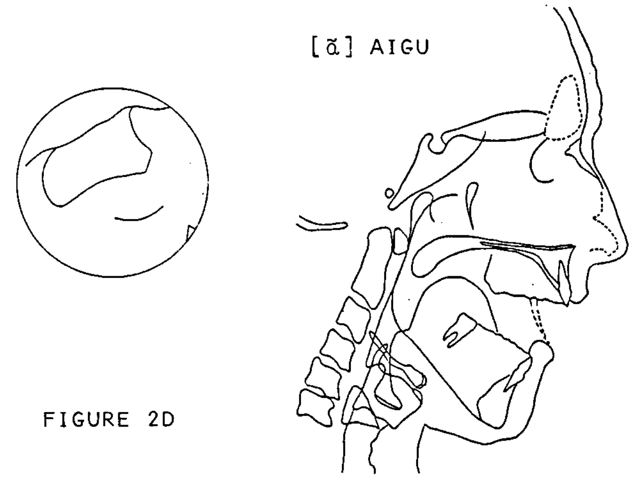


FIGURE 2D

endoscopiques de respiration.
Les résultats des mesures effectuées sur les calques radiographiques et endoscopiques sont consignés dans les tableaux I et II.

TABLEAU I : Mesures radiologiques.

	Relèvement	Fermeture
Respiration	0	7,5 mm
[i] GRAVE	7,5 mm	1 mm
[i] AIGU	4 mm	2 mm
[ã] GRAVE	2 mm	6 mm
[ã] AIGU	3 mm	4 mm

TABLEAU II : Mesures endoscopiques.

	Mouvements antéro-postérieurs	Mouvements latéraux
Respiration	21 mm	5,9 mm
[i] GRAVE	12,3 mm	1,7 mm
[i] AIGU	13,4 mm	3,7 mm
[ã] GRAVE	21 mm	5 mm
[ã] AIGU	14,5 mm	4 mm

IV. - COMMENTAIRES

Les tableaux I et II démontrent l'existence d'une corrélation positive entre le relèvement du voile et le rétrécissement des parois latérales du pharynx.
En revanche, relèvement et fermeture vélo-pharyngée sont inversement corrélés.
On a donc :
Relevé/Fermé/Rétréci ~ Abaissé/Ouvert/Elargi
ce qui signifie que la fermeture vélo-pharyngée ne peut être réalisée qu'en relevant le voile et en rétrécissant les parois latérales du pharynx. Or, dans le chant, le passage vélo-pharyngé et la totalité du pharynx doivent rester ouverts. Pour chanter on doit donc nécessairement abaisser le voile, ou du moins, utiliser des positions vélaires proches de l'abaissement. L'analyse détaillée des documents radiologiques et endoscopiques confirme d'ailleurs cette hypothèse.

A. - Voyelles parlées et voyelles chantées dans le grave :

Si le relèvement et l'accolement du voile sont toujours liés dans la parole, il n'en est pas de même dans le chant où soulèvement et accolement sont dissociés.
1. Pour la voyelle /i/ qui est la plus tendue et la plus fermée du système vocalique français, la radiographie du grave montre qu'il n'y a pas d'occlusion vélo-pharyngée comme dans la parole, mais au contraire, une ouverture de 1 mm.
L'endoscopie met en évidence un resserrement des parois latérales du pharynx beaucoup moins prononcé qu'en voix parlée et un relèvement du voile moins accentué (Fig. 2 A)
2. La réalisation de la voyelle nasale /ã/ nécessite aussi bien dans le registre grave que dans la parole une ouverture vélo-pharyngée plus importante.
Les calques radiologiques et endoscopiques montrent que pendant l'émission de la voyelle /ã/ dans le grave, la position du voile se rapproche de celle de la respiration (Fig. 2 B).

B. - Réalisations des voyelles dans l'aigu :

Lors de l'émission de l'aigu, le voile du palais occupe une position spécifique qu'il ne prend jamais dans la parole et qui est intermédiaire entre la position relevée/tendue des voyelles orales fermées et la position abaissée/détendue des voyelles nasales (Fig. 2 C et 2 D).
1. Pour la réalisation de la voyelle /i/, on observe sur les calques radiologiques une modification très nette de la forme du voile lorsque le sujet passe du grave à l'aigu (fig. 2 A et 2 C).
2. En revanche dans le cas de la voyelle /ã/ le voile du palais conserve sa forme mais change d'orientation pour l'aigu où il se redresse sans se soulever (Fig. 2 B et 2 D).

Dans le registre aigu, la différence entre orale et nasale tend à disparaître. Toutefois, pour essayer de maintenir la distinction des timbres vocaliques, le soprano a opéré de très légères modifications destinées à conserver les caractéristiques essentielles de ces deux types de voyelles : labialisation, augmentation de l'ouverture buccale, abaissement du voile, postériorisation de la masse linguale, dilatation de l'espace vélo-pharyngé pour /ã/ par rapport à /i/.
A ces modifications infimes, vient s'ajouter pour la voyelle nasale un rétrécissement important de l'espace compris entre la face endo-buccale du voile et le versant pharyngien de la langue, qui passe de 12 mm pour /i/ à 5 mm pour /ã/ (Fig. 2 C et 2 D). Le son vocalique étant moins oralisé va paraître davantage nasalisé (HUSSON /1/).
L'endoscopie ne fait pas apparaître de différences notables entre le /i/ et le /ã/ aigus. Par rapport à la voyelle orale, la nasale se caractérise par un très léger abaissement du voile associé à un infime élargissement des parois latérales du pharynx (Fig. 2 C et 2 D).

V. - DISCUSSION

A l'issue de cette étude, on peut se poser deux questions :
- pourquoi le voile du palais a-t-il cette position particulière dans l'aigu ?
- comment expliquer la sensation de soulèvement que ressentent les chanteurs alors que leur voile est dans une position relativement basse ?

1. La position spécifique du voile du palais dans l'aigu peut s'expliquer d'un point de vue physiologique et acoustique :

a) Du point de vue physiologique,
l'accolement maximal du voile contre la paroi postérieure du pharynx est associé pour la parole à un double effet de tension provenant
- des muscles du voile :
le péristaphylin interne (levator veli palatini) associé au palato-pharyngien et au palato-staphylin (musculus uvulae), et
- des muscles du pharynx :
le constricteur pharyngien supérieur qui agit sur les deux diamètres du pharynx (les constricteurs moyen et inférieur ayant surtout une action élévatrice).

Ces différents groupes musculaires fonctionnant en synergie, aboutissent à une fermeture de type sphinctérien qui risquerait d'entraîner un serrage au niveau du pharynx lors de l'émission de l'aigu. Pour éviter une trop grande contraction des muscles pharyngiens et vélaires, le chanteur va soulever son voile sans l'accoler à la paroi postérieure du pharynx grâce à l'action simultanée du péristaphylin interne et du palato-pharyngien (l'activité du palato-staphylin étant moindre dans ce cas).
Bien que son rôle ait été remis en question pour la parole, on peut envisager l'intervention, comme tenseur, du péristaphylin externe, seul muscle du voile totalement indépendant des autres muscles du pharynx.

b) Du point de vue acoustique et perceptif,
afin de mettre en évidence l'influence éventuelle de cette position du voile sur le timbre de la voix chantée, nous avons demandé à un soprano d'émettre des vocalises du grave à l'aigu, d'abord normalement, c'est-à-dire avec un voile en position basse, puis avec le voile du palais en position haute.
Dans le second cas, l'analyse acoustique fait apparaître une atténuation très nette du Singing Formant ainsi qu'une diminution importante de l'amplitude du vibrato de fréquence (qui passe de 11,58 Svt à 24,80 Svt, autrement dit de 1/4 de ton à un 1/2 ton).
Les deux vocalises soumises à l'appréciation de professionnels du chant ont été jugées à l'unanimité de la façon suivante : le son de la vocalise émise avec le voile du palais soulevé est qualifié de "plat", "terne", "écrasé"; alors que le son de la première vocalise émise avec le voile abaissé est considéré comme "beau", "rond" et "puissant".
Sans être en mesure d'affirmer l'existence d'une résonance nasale dans le chant, nous constatons avec TARNEAUD /2/ que l'accès de l'air pulmonaire aux fosses nasales "embellit le coloris de la voix".

2. Des générations de professeurs ont parlé de "la voûte du son" qu'il faut réaliser en "soulevant" le voile. De même, tous les chanteurs sont conscients de ce "soulèvement" qu'ils maîtrisent parfaitement. Or, ainsi que nous l'avons vu tout au long de cette étude, le voile du palais occupe dans l'aigu, une position relativement abaissée.
Afin de comprendre comment une telle erreur dans l'appréciation d'un mouvement pouvait être commise par des sujets dont les sensibilités internes sont extrêmement développées, nous avons réalisé un examen endoscopique complémentaire du voile et de la luette par voie buccale.
L'endoscopie exo- et endo-buccale nous a permis d'observer pour l'émission de l'aigu, une tension transversale du voile et un étirement des piliers postérieurs, associés à une hyper-rétraction de la luette.
C'est ce soulèvement important de la région uvulaire qui donne aux chanteurs l'illusion de soulever l'ensemble du voile; ce qui explique que cette erreur ait pu être perpétuée pendant si longtemps par les chanteurs et les professeurs de chant.

VI. - CONCLUSION

Dans la parole, lorsque le voile du palais est relevé au maximum, il vient s'accoler contre la paroi postérieure du pharynx et isole totalement le naso-pharynx de la cavité buccale; ce qui n'est jamais le cas dans le chant où la cavité nasale reste ouverte en permanence ainsi que le montrent les téléxéroradiographies que nous avons réalisées sur différents chanteurs professionnels.
Certains auteurs comme BARTHOLOMEW /3/ ou HUSLER /4/ estiment cette ouverture vélo-pharyngée indispensable dans le chant : "When singing, the nasal cavity must stay open (because it is) one of the main resonators in singing".
D'autres, comme BUNCH /5/ en nient l'existence : "Both actions (elevation and tension) are most important for speech and singing because this creates more resonating space in the oral pharynx and blocks off the nasal pharynx, preventing an undesirable nasal tone".
Sans prendre parti dans cette querelle, nous constatons néanmoins sur nos documents l'existence d'une ouverture vélo-pharyngée qui augmente du grave à l'aigu.
Quant à la position qu'occupe le voile du palais pour l'émission de l'aigu, nous pensons qu'elle résulte d'un équilibre physiologique entre les muscles vélaires et pharyngiens.
Des études complémentaires seront nécessaires pour préciser si cette position du voile n'est pas la conséquence d'autres coordinations motrices liées à l'augmentation de l'ouverture buccale et au relèvement de la base de la langue.

BIBLIOGRAPHIE

/1/. HUSSON, R., **La voix chantée**, Gauthier-Villars, Paris, 1960, 205 p.
/2/. TARNEAUD, J., **Le chant, sa construction, sa destruction**, Maloine, Paris, 1946, 135 p.
/3/. BARTHOLOMEW, W. T., "The role of imagery in voice teaching", **Proceedings of Music Teachers' National Association**, Oberlin, Ohio, 1936 : 78-93.
/4/. HUSLER, F. & RODD-MARLING, Y., **Singing, the physical nature of the vocal organ**, Hutchinson, London, 1976, 148 p.
/5/. BUNCH, M., **Dynamics of the singing voice**, Springer Verlag, Wien - New York, 1982, 156 p.
/6/. LEGENT, F., PERLEMUTER, L., VANDENBROUCK, C., **Cahiers d'Anatomie ORL**, vol. II : **Fosses nasales, pharynx**, Masson, Paris, 1986, 137 p.

THE PITCH OF GLIDE-LIKE F_0 CURVES IN VOTIC FOLK SONGS

JAAN ROSS

Dept. of Computational Linguistics
Institute of Language and Literature
Lauristini 6, 200 106 Tallinn
Estonia, USSR

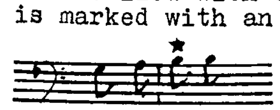
ABSTRACT

It has been shown recently that in Votic folk songs two successive tone glides, one rising and the other falling, create a clear unambiguous pitch sensation. An experiment was conducted where musically trained subjects had to match such F_0 curves with their voice. The stimuli consisted of four-note excerpts from the natural song performed by a low female voice, where the third note corresponded to a glide-like F_0 variation. The results confirm that the perception of such F_0 curves belongs to the category of fusion, which is characterized by the pitch sensation near the arithmetic mean of the terminal frequencies and a small dispersion of the subjects' responses. However, it seems more appropriate to define the pitch as F_0 at the point of two thirds of the overall duration of the note.

INTRODUCTION

In speech, permanent F_0 changes are very common while long time intervals with a stable pitch are perceived as something unusual. Music, on the contrary, is considered to be a sequence of notes, i.e. of discrete segments with a relatively stable pitch each. Actual relationship between pitch and frequency in musical performance can, however, be quite complex. When a particular note is perceived as having a stable pitch it does not necessarily mean that during this note F_0 should behave as stably as the pitch. Among such frequency changes which contribute more to the overall impression of sound than to the pitch sensation, the best known is vibrato. Vibrato can be defined as a simultaneous frequency, amplitude and phase modulation. Evidence can be found in literature that in European concert singing F_0 can be modulated within the limits of up to 2.5 semitones, the modulation rate being 5-8 Hz /1/.

Recently glide-like F_0 variations in notes with a stable pitch have been described in Votic one-voiced folk songs /2/. (The Votes are a Baltic-Finnic ethnic minority group living near Leningrad.) As was shown in /2/, glide-like F_0 variations in these songs tend to occur due to the coincidence of the metric accent, word stress and melody culmination. They have a shape somewhat similar to a circumflex where the right side is approximately twice as long as the left one. The difference between the initial and the final frequencies is considerable and in the case of the (left-side) rising glide can reach up to 4 semitones (see Fig. 1). As a matter of fact, the decision about the perceived pitch stability in such cases can be made without special experimental research. In ethnomusicology, a written transcription of performed song into conventional notation is an obligatory procedure and ethnomusicologists are very sensitive even to minor pitch changes which are beyond the limits of conventional notation. There is a whole arsenal of special symbols designed to mark down such changes. We have compared the conventional of the songs under investigation with the results of an F_0 extraction procedure and have not found any special symbols at the notes with a glide-like F_0 variation which means that these notes are perceived as having a stable pitch.
In this paper we consider glide-like F_0 variations mainly in one song called "Admonishment of the bride". This song belongs to the older layer of the Baltic-Finnic musical folklore, the origin of which is dated back at least as far as the 11th century. The song is characterized by a strongly recitative manner of singing and is performed by a low female voice in the frequency region of 180-250 Hz. During the song one and the same melody pattern is repeated 12 times with possible modifications. Glide-like F_0 variations occur most frequently where an odd repetition is passing over to an even one. The transition looks like this (the first note of the even repetition, which is always characterized with a glide-like F_0 pattern, is marked with an asterisk):



Such a transition does not occur at the end of even repetitions as they terminate a minor structural unit in the song. The total number of repetitions being twelve the present melodic segment occurs six times. The F_0 patterns of the notes marked with an asterisk are presented in Fig. 1. Fig. 2 displays the F_0 curves of the note next to that with an asterisk (i.e. the final note in the example presented above, or the 2nd note in the basic melody pattern). The difference between Figs. 1 and 2 is considerable and it seems necessary to find an answer to at least two questions in this connection. First, what is the F_0 value corresponding to the pitch of the note with an asterisk, and second, whether the pitch of the note with an asterisk is the same as the pitch of the next note. A psychoacoustical experiment was conducted in order to answer these questions.

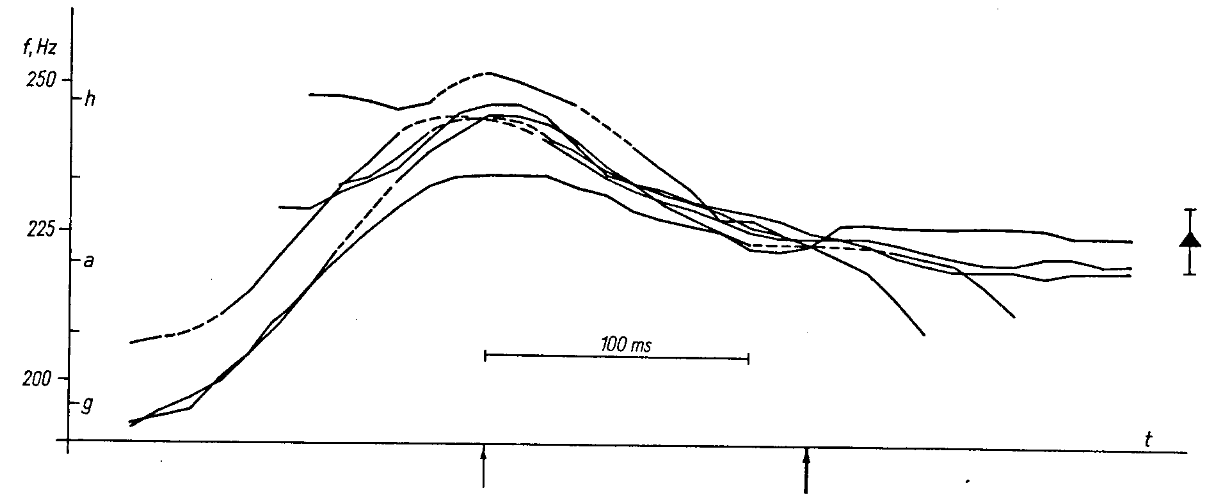


Fig. 1. F_0 patterns in a Votic wedding song which are perceived as having a stable unambiguous pitch. Six F_0 patterns are normalized with respect to the frequency maximum (marked with the left arrow) along the time axis (abscissa). Ordinate: the frequency scale (Hz) and its division according to the equal temperation. The right arrow on the abscissa roughly corresponds to the point at two thirds of the overall note duration. Mean F_0 values from a pitch-matching test, with standard deviation, are presented at the rightmost side.

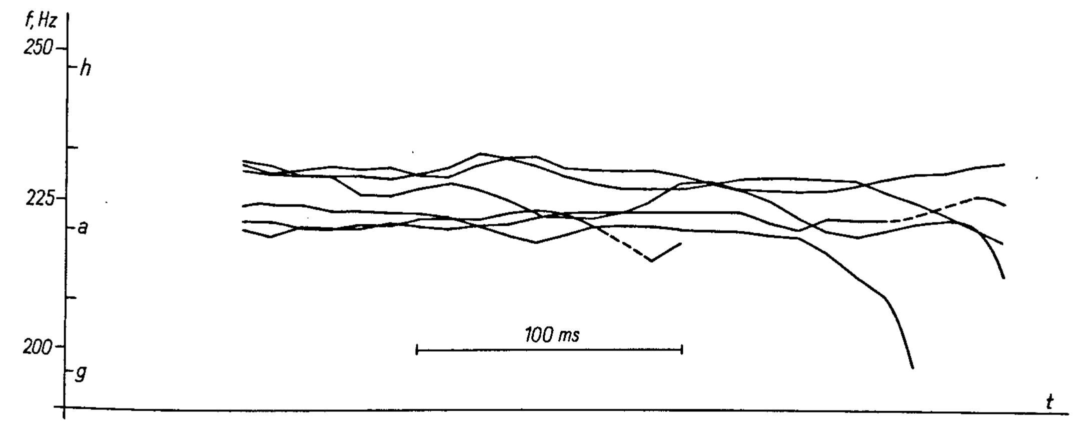


Fig. 2. F_0 patterns in a Votic wedding song which are perceived as having the same pitch as those in Fig. 1. Abscissa: time, ordinate: frequency scale (Hz) with its division according to the equal temperation.

PROCEDURE

The song "Admonishment of the bride" was stored on a magnetic disk of the EC1010 computer from a tape-recorder via a low-pass filter with a cutoff frequency of 5 kHz and a 12-bit A/D-converter with a sampling frequency of 10 kHz. Six odd-to-even transition samples were separated from the song and played back to subjects via a 12-bit D/A-converter, a lowpass filter with a cutoff frequency of 5 kHz and TDS-1 earphones. The subjects were free to regulate the SPL to a comfortable level and listen to every sample (almost identical to each other musically) as many times as they chose. The subjects were instructed to match the pitch of the third note in the melodic pattern with their voice. (This pattern was presented to them in a written form, too.)
Three subjects, two male and one female, participated in the experiment, all with a musical education at least on college level but with no reported absolute hearing. They were told to use octave transpositions of pitch if necessary in order to sing in a more comfortable frequency range. As all 3 subjects responded to 6 melodic patterns, the total number of answers was 18. The answers were then stored on a magnetic disk using the hardware mentioned above and run through a pitch detection program. The resulting F_o response curves of all three subjects were smooth enough with a frequency change of no more than 5 per cent. The frequency which occurred most often was chosen as the response in every case.

RESULTS AND DISCUSSION

Both male subjects were able to match the pitch of the stimuli with no octave transpositions, i.e. they used the same frequency range as the singer originally did. The female subject performed the task one octave higher. So, the frequency values of her responses were replaced by twice smaller ones during the following computation.
We assumed that in the original song the singer uses a kind of internal tone scale and frequency deviations in the case of one and the same note may occur only incidentally. The mean F_o value of all of the 18 responses are presented on the right side of Fig. 1, with standard deviation. As one can see, the range of standard deviation is practically identical to the range of the F_o variation in Fig. 2. So we can conclude that the pitches of two successive h's (3rd and 4th note in the note example above) are equal despite the remarkable differences in the corresponding F_o patterns.
Nábělek, Nábělek and Hirsh /3/ have studied the pitch of sound bursts where two terminal frequencies are connected by a linear frequency change. They have found that three modes of perception can be distinguished with such glides, namely fusion, separation and indecision. Fusion is characterized by a pitch sensation near the arithmetic mean of the terminal frequencies and a small dispersion of responses. In the case of separation the pitch of the glide is matched at one or both terminal frequencies. Indecision means that the subjects' responses are widely distributed among the frequency axis between the terminal frequencies. Glide-like F_o patterns in Votic songs seem to belong to the category of fusion, as the dispersion of the subjects' responses is small and the perceived pitch corresponds neither to the initial nor to the final frequency but is rather matched at the middle of the glide. Nábělek et al. have concluded that in the case of tone bursts with a linear change of frequency during 100 per cent of the burst duration, fusion occurs when the difference between terminal frequencies Δf is not too great and the duration of the glide t_b not too long. As one can see in Fig. 1, in Votic songs $\Delta f \simeq 50$ Hz and $t_b \simeq 100$ ms in the case of a rising glide (the left part of the circumflex) and $\Delta f \simeq 30$ Hz and $t_b \simeq 200$ ms in the case of a falling glide (the right part of the circumflex). When we compare the values of Δf and t_b with those of Nábělek et al. considering the left and the right sides of the F_o pattern separately, we find that single tone bursts with such parameters clearly fall to the category of fusion. So we can conclude that the results are in good agreement with their investigation.
However, the conclusions by Nábělek et al. are rather general in nature as they enable us only roughly to determine the F_o value which corresponds to the perceived pitch. In their following work /4/ the authors wrote even more cautiously that "in fusion ... the sound burst was matched by a single frequency located somewhere between the initial and final frequencies". In the experiment of Lublinskaya /5/ subjects had to mimic with their voice the F_o variation in synthetic vowels /a/ with a glide-like frequency change. The author proposes that the most important task is to determine the initial and final frequencies of a glide whereas the pitch as corresponding to their arithmetic mean can be considered their derivative. The comparison of the original F_o and the response curves in her study confirm that it is the terminal frequencies that are stressed in subjects' responses.
Rossi /6/ has shown that the pitch of the synthetic vowel /a/ with both an ascending and a descending glide-like F_o change ($\Delta f = 23$ Hz, $t_b = 200$ ms) is perceived at the point of two thirds of the overall duration of the vowel. This point approximately corresponds to that marked with an arrow at the right side of Fig. 1. As one can see, at this point all six F_o curves are passing quite a narrow frequency region of 224-226 Hz which very closely corresponds to the mean of the subjects' responses in the described pitch-matching experiment. So our experimental results well agree with those of Rossi, too.

CONCLUSIONS

The pitch of tone glides has been explained on the basis of hypothetic mechanisms operating both in frequency and time domain. When pitch is measured in the frequency domain, it is the determination of the initial and final frequencies that seems to be of the greatest importance. In the time domain, the pitch should correspond to the F_o value at the point of two thirds of the overall duration of the stimulus. The results of the present experiment can be explained both on the basis of the frequency and the time domain hypotheses, but as the latter allows us to determine the F_o value more precisely, this explanation should be preferred.

ACKNOWLEDGEMENTS

Thanks are due to Ingrid Rüütel for valuable ethnomusicological comments, and Sirje Murumets for editorial assistance.

REFERENCES

/I/ В.П.Морозов. Биофизические основы вокальной речи. - Л.: Наука, I977.
/2/ I.Rüütel, J.Ross. A study of pitch contours and the scale structure in Votic folk music. - Preprint KKI-37. - Tallinn, 1985.
/3/ I.V.Nábělek, A.K.Nábělek, I.J.Hirsh. Pitch of tone bursts of changing frequency // J. Acoust. Soc. Am. - Vol. 48. - 1970. - P. 536-553.
/4/ I.V.Nábělek, A.K.Nábělek, I.J.Hirsh. Pitch of sound bursts with continuous or discontinuous change of frequency // J. Acoust. Soc. Am. - Vol. 53. - 1973. - P. 1305-1312.
/5/ В.В.Люблинская. Воспроизведение голосом изменения частоты основного тона звука // Сенсорные системы / Вопросы теории и методов исследования восприятия речевых сигналов. - № 2. - Л., I97I. - С. 92-98.

/6/ M.Rossi. La perception des glissandos descedants dans les contours prosodiques // Phonetica. - Vol. 35. - 1978. - P. 11-40.

APPLICATION OF AUTOMATED IDENTIFICATION METHODS OF BOW STROKES TO MUSICAL FOLKLORE RESEARCH

Dedicated to Professor
Jadvyga Čiurlionytė

Artūras Medonis
Academy of Sciences
of the Lithuanian SSR
Vilnius, USSR

Birutė Sinkevičiūtė
Dept. of String Instruments of the
Lithuanian SSR State Conservatoire
Vilnius, USSR

ABSTRACT

The purpose of the presented paper is to search for the common and distinctive features of an automated investigation of music/speech and simple speech signals. An algorithm of musical parameter estimation is based on the application of speech parameter recognition. Topical aspects of the automated ciphering in case the height of the musical folklore sounds is the same are analysed in our report.

INTRODUCTION

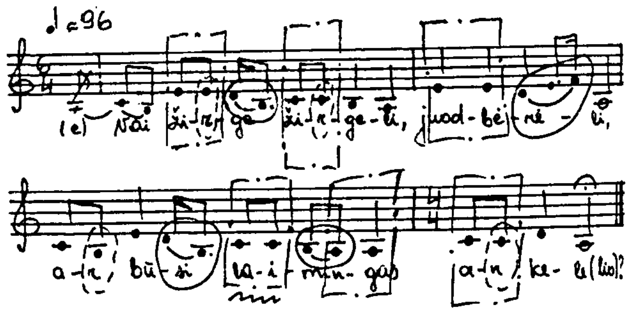
"Lithuanian people's songs art is exceptionally rich and various. The Lithuanian folklore Manuscript fund stores about half a million records of songs"./7 /

The most urgent problems of the Lithuanianfolklore songs were reviewed by Professor J.Čiurlionytė for the first time outside the republic /9, 10/. After that the songs are investigated and systematized in the following organizations:
- the Folk Music Study of the Lithuanian State Conservatoire (founded by Professor J.Čiurlionytė)/6, 8/,
- the Folklore Department of the Institute of the Lithuanian Language and Literature of the Lithuanian Academy of Sciences /7, 17/.

An evolution of folklore witnesses the improvement of folklore investigation methods was attained thanks to the application of new technical equipment (phonograph, tape-records). Lately folklorists often solve their problems with the help of computers, i.e. cipher, analyse and systematize the songs.

From the first stages in folklore investigations both the linguists and etnomusicologists are excited by these problems in the junction of the folklore music and linguistic text / 1,11, 17/. E.g., Vice-president of the Folklore Board of the USSR Composer's Union Alexeyev E. explores the interrelation between musical intonation and verbal inflection on the basis of Lithuanian folk-songs /1/.

Our report deals with a specific aspect of the above junction: it presents an estimation algorithm of musical durations; the algorithm is based on the application of characteristic speech parameters when the height of the sounds analysed is the same (Note example I, case /a/).



Note example I.
/a/ - a case when the height of the sounds sung is the same;
/b/ - consonants with independent musical height and duration;
/c/ - a melodic phrase performed by one and the same vowel;
/d/ - a case when both vowels of diphtong have independent musical durations.

The purpose of the presented paper is to discuss the common and distinctive features of automated investigation of music/speech and simple speech signals.

The 1986 International Computer Music Conference (ICMC) was held in Hague, the Netherland. October 20-24. We presented a report in which we discussed our main topic "On the identification of violin strokes in a real-time performance system" in this ICMC. Now let us review the report in short and to present our main material.

METHOD

Our paper /18/ deals with the automated identification of violin strokes in case the height of the sounds played is the same. E.g. it happens in the main theme of Concert No 2, E-dur by J.S.Bach (for violin and chamber orchestra).Estimations by the statistics

$$L_k=\left(\sum_{i=1}^{N} x_i^2/N\right)^{1/2}, \quad k=1,2,\ldots \qquad (1)$$

are obviously sufficient for sounds and pauses in a pizzicato case but it is not sufficient to identify martele strokes: (1) the estimates slightly differ from one another and the corresponding algorithm identifies the beginning and the end of the strokes unprecisely.

The violin plate bears a property to resonate the sounds played and to cease ringing gradually after stopping playing. Therefore a physical interpretation of the martele envelope is the following: at the beginning of a new stroke we listen to the sounds of new and earlier strokes resonated by a violin plate.

Analogous features of merging sounds are typical of speech analysis too: the signal is segmented in order to determine the limits of separate sounds. Segmentation of diphtongs is especially difficult as the Lithuanian linguist A.Pakerys notes / 15/.

We proposed not only to measure the sound parameters, but also to register the supplementary information obtained from a digital bow /4/.

Formalization of the above statement: let us consider a violin stroke identification function $F(y)$, where $y=\varphi(\nu,L,t)$ - is the function of violin sound determination; ν, L, t are the violin sound parameters: pitch, intensity and duration, correspondingly. The aim of the automated identification is to define the moments t_j:

$$F(y)=\begin{cases} a, & t<t_j \quad (a \leftrightarrow \sqcap\text{-}m.n.^{*}) \\ b, & t>t_j \quad (b \leftrightarrow V\text{-}m.n.) \end{cases} \quad \text{in case } \nu=const$$

The segments $X_{K1}, X_{K2}, \ldots, X_{KN} : L_{K+1} \gg L_K$ ($k=1,2,\ldots$) were investigated for this purpose. The algorithm is sufficient for pizzicato stroke identification, but in a martele case the segments $k: L_{K+1} \gg L_K$ were not detected. Therefore, we proposed to consider the function $F(y,z)$, where $y=\varphi(\nu,L,t)$ and $z=\psi(t_i)$:

$$\gamma(t_i)=\begin{cases} 1, & \text{as } \psi'(t_i)<\psi'(t_{i+1}) \quad (\sqcap\text{-}m.n.) \\ 0, & \text{as } \psi'(t_i)=0 \text{ or } \psi'(t_i) \nexists \\ -1 & \text{as } \psi'(t_i)>\psi'(t_{i+1}) \quad (V\text{-}m.n.) \end{cases} \qquad (2)$$

Now we deal with the application of the above methods to the automation of Lithuanian musical folklore analysis. Automated ciphering of the original, in case the height of the sounds sung is the same, (note example I, case /a/) is an urgent problem. Case /b/ of the note example illustrates the situation when a consonant is in good agreement with musical duration.

* m.n. - in musical notation

A phonetic syllable in the speech analysis is determined as the least segment of speech torrent, unit of the pronunciation, which forms the words rhythmically and with emphasis/14/. In literature we did not notise a monosemantical definition of a phonetic syllable in the aspect of musical folklore analysis; let us denote it in terms of a "musical syllable".

There are two points of view between etnomusicologists on the above problem:
(a) musical syllable is not a structural sound which has no influence on the definition of etnomusicological parameters of the investigated melody: metre-rhythm structure, ambitus, the stable marginal sounds of the melodic vertical - the lower and upper tonic - etc.
(b) musical syllable is namely a structural sound which is an independent unit and therefore is fixed in musical notation.

There are many peculiarities in folklore singing and untraditional elements of musical notation, used to express them, e.g. etc.

Let us consider a musical syllable duration (MSD) function $F(y)$, where $y=\varphi(\nu,L,t)$ is a function of a musical sound determination. The aim of the MSD identification is to define the moments t_j:

$$F(y)=\begin{cases} a, & t<t_j \\ b, & t>t_j \end{cases}, \quad \text{when } \nu=const$$

For this aim the segments $k: L_{K+1} \gg L_K$ are searched as they are presented in the methods for violin stroke identification. The algorithm is not always sufficient (e.g. sound intensity variation is often possible in the same vowel, as it is shown in fig2 for vowel "ė"). Supplement parameter has to be applied to musical parameter determination: classification parameter "voiced speech/unvoiced speech segments" are widely used in speech analysis / e.g.16/.Conseguently it is expedient to consider the function $F(y,z)$: $y=\varphi(\nu,L,t)$, $z=\psi(\tau)$ and

$$\psi(\tau)=\begin{cases} 1, & \text{when } \tau \text{ is a voiced speech segment} \\ 0, & \text{when } \tau \text{ is an unvoiced speech segment.} \end{cases} \qquad (3)$$

The value of parameter τ can estimated using one of the numerous algorithms discussed in special literature /16, 19/.

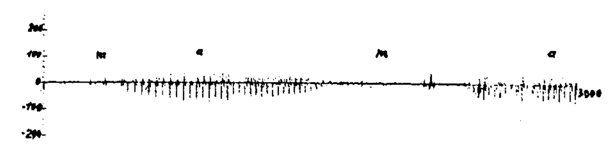


Fig.1 Acoustic signal of word sung " mama "

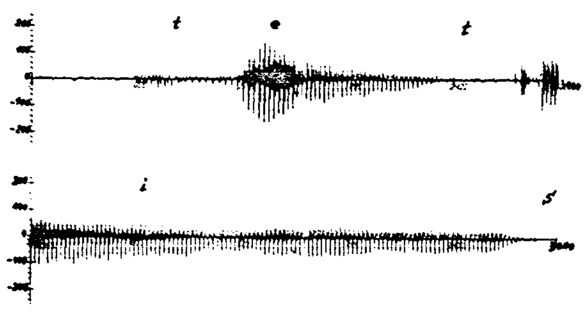
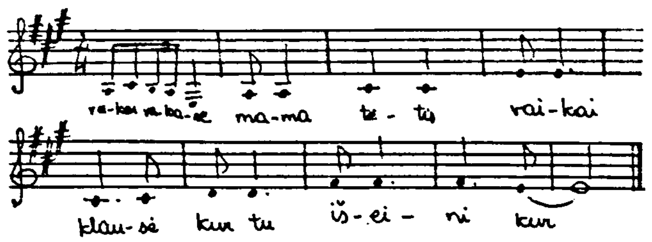


Fig.2 Acoustic signal of word sung "tėtis"

RESULTS

A musical fragment was introduced into the universal computer BESM-6 (Note example II). The sound waveform was sampled at 4 kHz because the fragment was a low register melody. For 100 ms acoustical segments the intensity was computed according to statistics (1) and an algorithm of a delay function $D(p)$ was applied to estimate the pitch. The estimates of ν and L were not sufficient for the definition of musical duration. Therefore, it is expedient to deal with the function with more parameters, i,e, to describe $F(y,z)$, where $z=\psi(\tau)$ .Estimation of the algorithm for segment vocalization is based on a common interpretation of energy and zero crossed frequency functions. An application of this methods gives more concrete results in comparison with the values of $F(y)$.



Note example II.

Pitch (Hz)	122.56	125.41	124.03	132.32	133.85	
Intensity	19.741	20.319	24.451	8.572	6.748	
Sound	-a-	-a-	-a-	-m-	-m-	

Pitch (Hz)	138.33	135.51	129.03	131.41	132.11
Intensity	17.642	20.809	17.642	19.581	20.499
Sound	-a-	-a-	-a-	-a-	-a

Table I. Estimates of parameters and of the sung word "mama"

x x x

In comparison with a real-time performance system it is sufficient to work in interactive regime when the musical folklore signal is processed. That allows to widely use spectral algorithms for pitch determination. In contrast, violin stroke identification requires the application of fast algorithms. One of such is presented in our paper /18/. It reflects an effective utilization of a delay function

$$D(p)=\sum_{k=1}^{M}|y_k-y_{k+p}|,\quad M<N,\quad p=0,1,\dots,N-M$$

As usual, in speech signal processing one of the three models is applied: excitation model, perception model or mathematical model/12/.

By extending the conclusion of Hess our point of view consists in that the application of the known pure mathematical methods is not sufficient in music/speech signal processing. Perception models, based on the musical knowledge, are preferable. The results below of our experiment done on the pitch determination illustrate this standpoint rather well.

Let us consider the following mathematical model of pitch determination

$$y(t)=f(t)+\xi(t),\quad 0\le t<\infty \qquad (4)$$

where $f(t)=\sum_{j=1}^{\infty}A_j\cdot\sin(2\pi\cdot\nu\cdot j\cdot t+\varphi_j)$

$y(t+T)=y(t)$, T is a period, $\nu=1/T$ is a pitch of the function $y(t)$; $\xi(t)$ is a stationary random sequence in a wide-sense. The aim is to estimate a parameter $\hat{\nu}=\hat{\nu}(y_1,\dots,y_N)$. We applied the following algorithms of spectral analysis:

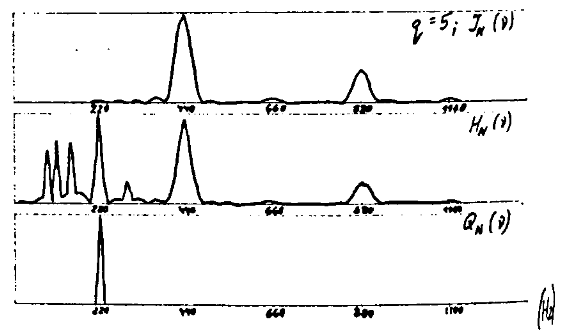


Fig.3

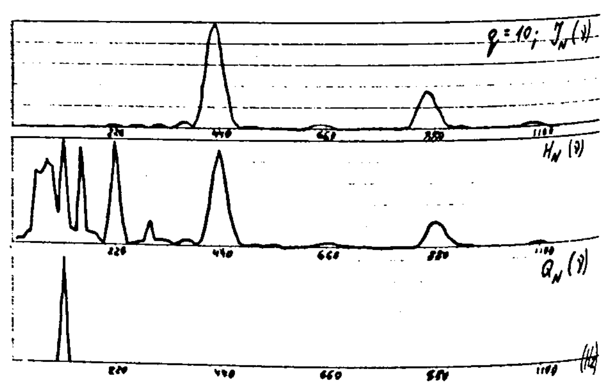


Fig.4

Fig.3, 4 illustrate graphic represantations of the functions $I_N(\nu)$, $H_N(\nu)$, $Q_N(\nu)$ obtained by theoretical series (8).

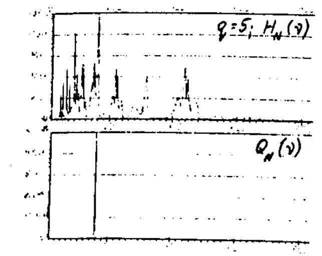


Fig.5

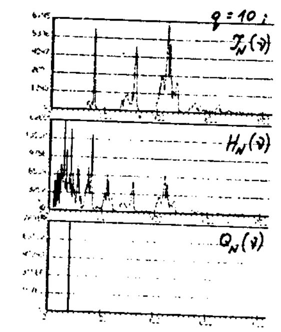


Fig.6

Fig.5, 6 illustrate graphic represantations of $I_N(\nu)$, $H_N(\nu)$, $Q_N(\nu)$ generated by a real series of sound "ė"

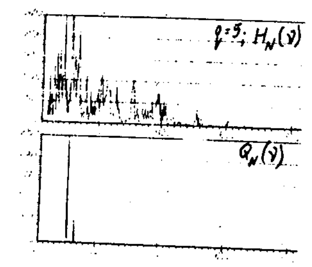


Fig.7

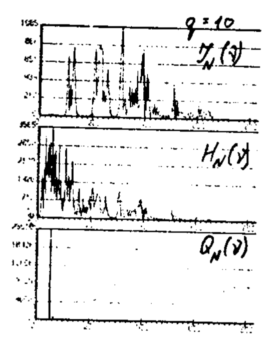


Fig.8

Fig.7, 8 illustrate graphic represantations of $I_N(\nu)$, $H_N(\nu)$, $Q_N(\nu)$ generated by a real series of sound "a"

I. $$I_N(\lambda)=\frac{\Delta t}{2\pi N}\left|\sum_{k=1}^{N}(y_k-\bar{y})\cdot e^{-ik\cdot\Delta t\cdot\lambda}\right|^2,\quad \lambda\ge 0:\ \lambda=2\pi\nu \qquad (5)$$

is a periodogram of time series (4) at the points $\lambda_j=\frac{2\pi}{N'\cdot\Delta t}\cdot j$, where $j=0,1,\dots,N'-1$; $N'=2^{\ell}\ge N$ is a selective number by a user. The values of the periodogram are calculated on the basis of the algorithm of FFT.

$\hat{\nu}=\Delta t\cdot[\arg\max_{\nu} I_N(2\pi\nu)]$ -the pitch estimate (5') by the above algorithm

II. $H_N(\nu)=\sum_{j=1}^{q}I_N(2\pi\cdot\nu\cdot j)$

-the sum of periodograms (6)

$\hat{\nu}=\Delta t\cdot[\arg\max_{\nu} H_N(\nu)]$

-the pitch estimate (6')

III. $Q_N(\nu)=\prod_{j=1}^{q}I_N(2\pi\cdot\nu\cdot j)$

-the product of periodograms (7)

$\hat{\nu}=\Delta t\cdot[\arg\max_{\nu} Q_N(\nu)]$

-the pitch estimate (7')

We deal with one theoretical and two real time series. The modelling series is

$$y_k=\sum_{j=1}^{5}A_j\cdot\sin(2\pi\cdot\nu\cdot j\cdot k\cdot\Delta t)+\varepsilon_k,\quad k=1,2,\dots,N \qquad (8)$$

The meanings of parameters are the following $N=2^{\ell}$, $\ell=10$, $\Delta t=0.02$ (MS)

$\nu_N=1/(2\cdot\Delta t)=25$ (kHz) the Naikvist's frequency

$\nu=220$ Hz; $A_1=0.1$, $A_2=1.0$, $A_3=0.2$, $A_4=0.6$, $A_5=0.15$

$\varepsilon_k\in N(0,\sigma^2)$ where $\sigma=0.1$, $q=5$ and $q=10$.

The results of application of algorithms (5)-(7) are shown in Fig.3-4. Pitch estimate $\hat{\nu}=220$ Hz corresponds to the given value ν in case $q=5$. In another case, as $q=10$, pitch estimate $\hat{\nu}=110$ Hz. The estimate is possible in a mathematical algorithm sense, but it is not logical in the aspect of musical theory. As it is well known /13/ in the theory, harmonics form consistent series of musical intervals: octava, quinta, quarta, b.tertia etc. There are following relations of harmonic frequencies in our example, as $q=10$: 220/110=2/1 -octava, 440/220=2/1 -octava, 660/440=3/2 -quinta etc. Octava interval is repeated in our series, but it is impossible in the aspect of musical theory (as untertone does not exist in a musical sounds). Pitch estimate $\hat{\nu}=220$ Hz generates necessary series of musical intervals,therefore, it is true. This musical knowledge is laid in perception models, based on the application of algorithms of harmonic sieve type/2 /.

Real time series of the sounds "a" and "ė" corresponds to the words "mama", "tėtis" of a music/speech signal. Selection of a parameter q has the influence on the estimation result, as shown in Fig.(5)-(8). Therefore, the application of perception models is preferable again.

ACKNOWLEDGMENTS

We highly appreciate the attention of our colleagues: V.Baranauskienė, K.Augutis, T. Juzeliūnas, A.Dikčius (the Lithuanian State Conservatoire), V.Pikelis and J.Kazlauskaitė (Institute of Mathematics and Cybernetics of the Lithuanian Academy of Sciences).

REFERENCES

1. Alexeyev E.(1986). The Pitch Nature of Primitive Singing. Sovetsky Kompozitor, Moscow, -240p. (in Russian).
2. Allik J., Mikhla M., Ross J.(1984).Comment on "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception".JASA 75, p.1855-1857.
3. Ambrazas A.(1981). New methods for cipheringand catalogation of folklore melodies. Music, ISSUE 1, p.121-123(in Lith.).
4. Askenfelt A.(1984). Study your bowing technique! -Quarterly Progress and Status Report, No 1, Dept of Speech Communication and Music Acoustics, Royal Institute of Technology, Stockholm.
5. There come nine-corn reindeer running. (1986). Folklore collection. Vilnius, 431p
6. Baranauskienė V., Medonis A., Medonis E, Sinkevičiūtė B.(1986). Some algorithms of Lithuanian folklore unanimous songs (in Lithuanian, to be printed).
7. Collection of the Lithuanian songs(1980) vol.I,The Songs of children. Prepared by the Lithuanian SSR Academy of Sciences, Vilnius, Vaga, -760p.
8. Četkauskaitė G.(1981). The Dzūkai Melodies, Vilnius, Vaga, -621p.(in Lith.).
9. Čiurlionytė J.(1949). Lithuanian folksong, Sovetskaya Muzika, No 6, 60-64p. (in Russian).
10. Čiurlionytė J.(1955).From the history

of the Lithuanian folk-song. Sovetskaya Muzika, No 2, 69-77p. (in Russian).
11. Čiurlionytė J.(1984). A song as the research object. Collection of papers "Daily works", Vilnius, Vaga, 46-53p.(in Lith).
12. Hess W.(1983). Pitch determination of speech signals. Berlin, Springer, -701p.
13. Kazlauskas J.(1975). Musical instruments and score. Vilnius, Vaga,-222p. (in Lithuanian).
14. Lithuanian Dictionary(1981). vol XIII, Vilnius, Mokslas, p.872-873 (in lithuan.)
15. Pakerys A.(1982). The Prosody of the standard Lithuanian language, Vilnius, Mokslas, -214p.(in Lithuanian).
16. Rabiner L., Shafer R.(1981). Digital processing of speech signals.
17. Sauka D.(1986). Lithuanian folklore, Vilnius, Vaga, -317p.(in Russian).
18. Sinkevičiūtė B., Sondeckis S., Medonis A.(1986). On the identification of violin strokes in a real-time performance system. Proceedings of the XII ICMC, 187-191p.
19. Zue V.(1985). The Use of Speech knowledge in Automatic Speech recognition. Proceedings of the IEEE, v.73, No 11,p75-91

APPENDIX (CONCLUSIONS)

The main points of the report are:
1) discussion of the problem of a searching for the common and distinctive features of an automated investigation of music/speech and simple speech signals on the basis of Lithuanian musical folklore (see note ex I)
2) parameters of the musical and speech processing are closely interconnected; therefore their common interpretation makes it possible to estimate musical duration of sounds more precisely in case their height is the same;
3) an exact definition of a term "musical syllable" is a topical problem in the aspect of automated ciphering of melodies. There are same specific features of music/speech sounds: sometimes consonants are in agreement with independent musical height and duration (note example I), necessary to express; another well-known case deals with the singing of some melodic fragment performed by one and the same vowel. Both the the specific elements are distinctive features of an automated investigation of music/speech and simple speech signals;
4) there are some common features of the above signals:
   a) acoustic parameter - pitch, intensity and duration - are most importance for both types of signals,
   b) a necessity of diphtong segmentation in music/speech and simple speech processing (Note ex.I /d/);
5) direct application of pure mathematical methods are not sufficient in music/speech signal processing.

ON THE PHONOLOGY OF YAKUT FOLK CHANTS

Aleksej Klyuchevsky

Centre of Studies in Musical
Folklore of Siberia and the Far East
Novosibirsk Conservatoire
Novosibirsk, USSR, 630099

Yurij Sheikin

Dept. of the Philology of the
North, Institute of Language,
Literature and History
Yakutsk, Yakut ASSR
USSR, 677000

ABSTRACT

This is an attempt at applying the linguistic phonetic approach to the analysis of chants /1/. A new phonic notation is suggested and peculiarities of the phonological system of Yakut chants are discussed.

INTRODUCTION

Musical culture may be seen as one more level of natural language with its own paradigmatics and syntagmatics and is in some special way connected with the phonological level of language, above all, to prosody. Or, to use another terminology, musical culture lies within the linguistic competence. Chanting brings into operation some additional rules affecting all the components of grammar and imposes some special markings in the lexicon. The semantic representation also changes. This hinders the understanding of a chant by the language-users whose linguistic competence does not include musical culture. In this case non-musical speech may be regarded as an unmarked performance with a zero feature.

1.0. The traditional five-lined tempered octave system, adequate to the European culture as it has evolved, cannot efficiently express the system of meaningful musical stylistic characteristics (musical space and time) of non-European cultures. That is why musical structures that are felt to be very simple appear rather awkward when it comes to notation. Every musical system has its own way of segmenting the sonic ambit (the scope within which melodic development occurs). Proceeding from the idea of the discrete character of modal cells, we suggest here a music notation which stands in the same relation to traditional notation in music as phonological notation is to phonetic transcription.

1.1. Basic notions.

1.1.1. Basic tones (bases). A folk chant is oriented around several basic tones (or bases), that is, pitch constants structuring musical and intonational development. Each basic tone (base) may be represented by one pitch or by a pitch zone, depending on the structural peculiarities of the melody. This allows us to classify chants according to the types of the development of basic tones (measured in kHz), as we do in classifying intonational patterns in sentences.

1.1.2. Musical Intonemes. Elementary units of a chant are discrete modal cells which we shall call musical intonemes, or, simply, intonemes.

Intonemes are distinguished according to the following characteristics:
1) basic or gliding, that is, oriented or not oriented towards basic tones.

Basic intonemes differ depending on the basic tone they are oriented around. Here they are divided into upper and lower.

Gliding intonemes are oriented around basic intonemes and are divided into falling and rising. In our text gliding intonemes do not occur.
2) simple - complex. Simple intonemes can be short or long. Short intonemes include those with a duration of 1, 2 or 3 conventional units.

Intonemes that are longer than 4 units are called long.

Complex intonemes consist of several simple intonemes linked together.
3) ornamented - unornamented. Ornamentation is a slight constriction of the upper part of the vocal cords and a uvular trill, an active sound ejection developing into a normative constant which, with Yakuts, goes beyond speech phonation. Ornamentation can be partial preceeding, partial following and continuous.

Such ornamentation may be quite prolonged at a certain pitch. In this case, it forms a characteristic tone at a stable pitch, sustained above the fundamental tone, which gives the effect of a binary dependent phonation. This is a special type of Yakut chanting, called kïlïhaq.

Intonemes may be preceded or followed by a checked intoneme, that is, an intoneme with a duration of less than 1 unit.

1.2. Notation symbols and intoneme combinations.

(The present notation is slightly suggestive of the generally accepted notation in music, of the staves in particular. We think it quite possible that a more consistently symbolic notation /say, linear/ might be proposed. The process of analysis itself will determine the most suitable notation).
1). Basic tone (bases) ______ measured in Hz

2) a short simple unornamented basic lower intoneme (duration - 1)
3) , short simple unornamented basic intonemes (duration-2 and 3 respectively).
4) a long simple unornamented basic intonemes (duration - more than 4)
5) preceding ornamentation
following ornamentation
continuous ornamentation
6) a basic intoneme with a preceding checked element, and
a basic intoneme with a following checked element
7) a rising unornamented gliding intoneme (duration - 1)
a falling unornamented gliding intoneme (duration - 1)
8) Prolongation, length, ornamentation and checked elements for gliding intonemes have the same symbols as for basic intonemes.
For instance,

The elements, described may occur in various combinations. The link between elements is shown by juxtaposition or by a connecting line. E.g.

1.3. Groupes of intonemes possess certain prosodic features (the type of prosody is clear from the context), and there are specific peculiarities of combinations of the prosodic features and units. A prosodic feature symbol is placed in front of brackets. In case of hierachically ordered prosodic features the order of the feature is shown by its position with respect to the brackets.
1) кIlIhaq (see above) + kIl ( ).
In кIlIhaq bracketed are only unornamented intonemes.
2) Inhalatory chanting (inspiration) - + inspir. Long ornamented intonemes do not occur in inhalatory chanting.
3) Exhalatory chanting (expiration) is unmarked Yet, when alternating with inhalatory chanting, it may be designated as + expir.
However, this seems, strictly speaking, redundant.
4) Continuous and simultaneous pharyngal and laringal constriction - phar./lar. ( )
Within brackets all elements may occur.
The above elements, however, do not exhaust the vast variety of prosodic features.

2.2. The beginning of the chant (to denote it we shall use the term musical initial, hereafter initial) sets the articulary position of the chant, in other words, the pitch relations to be fulfilled in chanting. That is to say, it determines the basic tones (bases). They are given by simple vowels, usually of the simple vowel triangle. In our example, though, the vowel I is introduced there, too. Every syllable of the initial opens with prothetic or epenthetic consonants, conditioned by following vowels. The initial of the chant in question is hIn-ǰe-buo-a (I-e-u-o-a).
Examples of other initials: šie-ñej-ñeje (i-e-e-i-e-i-e), ijau-ijau-ijau (i-a-u-i-a-u) (Lamut chants), jawa-jawa (i-a-u-i-a-u) (Udihe chants).

3.0. In Yakut folk chants we observe a number of phonetic peculiarities partially shared by all texts of poetic folklore.
Yakut vowels may be described in terms of the following binary distinctive features: high, low, back, labialized, long (tense?).
Due to vowel harmony the number of distinctive features to oppose the vowels within one wordform is reduced. This allows us to regard vowel harmony as a prosodic feature. The hierachy of vowel harmonies may be shown by the position with respect to the brackets (analogous to 1.3.)
In the Yakut language there is a consistent palatal harmony and a more or less consistent labial harmony /2/. The degree of aperture is also harmonized but we shall not discuss it here. The zone (± back) and rounding (± lab.) features are indicated in front of the brackets, whereas the indications as to the rise and quantity of vowels are given inside. So, inside, we find a kind of phonological vowel matrix where the zone and rounding qualifications are not made explicit. To designate them we shall use capital letters denoting back non-labial vowels.
3.1. There are lexemes where labial or palatal harmony is irrelevant. E.g.: doydu-daydI (the world) :±lab.+ back (dAydI), sirey-sIray (a face) : - lab. ±back (sIray).
This freedom, like prosodic features of other languages, is used in poetry in two ways: first, in alliterations, second, together with alliterations, in semantic characterization, since in many attributes of the upper world of the Yakut mythological cosmos we find lexemes with + lab. - back features, whereas lexemes with - lab. + back features /3/ are found to denote attributes of the lower world, that is, their full distinctive-feature (and, thus, articulatory) opposition is made effective use of.

3.2. Consonants also show a tendency toward a distinctive-feature homogeneity. Also used as a poetic means is an optional distant consonant assimilation accoding to nasality feature, that is, a kind of nasal harmony. E.g.: maɣan - maŋan - white, baɣana - maɣana -maŋana -a post supporting the yurta.

3.3. In chanting the opposition with respect to the quantity of vowels becomes irrelevant, since short vowels may be drawled, too.

3.4. Drawling in a chant may bring about the neutralization with respect to the harmony of wordform elements or separate vowels which are intoneme-carriers. For instance, a non-labial back dyphthong Ia is monophthongized and becomes a neutral vowel.

3.5. We have also observed certain dissimilation processes at junctures of wordforms within syntagmas which become phonetic words.

E.g.: kőrönnör → köröndör - looking;
elbeq ügüs kemŋe → el-be-ǰü-üos-kem-ŋe - much time.

3.6. In drawling and tone-falling narrow vowels are dyphthongized. E.g.: oɣotun - oɣotuon - child; körbötün - körbötüön - (he, she) will look.

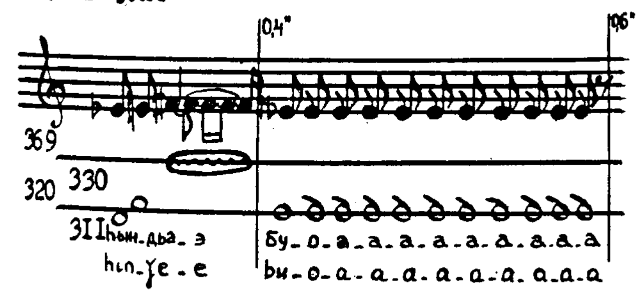
NOTES & REFERENCES

/1/ We chose a text in the Yakut language since due to the regularity of the phonological and grammatic structures Yakut texts are most convenient for all types of linguistic analysis. We hope to apply this approach to the analysis of other Syberian folk material.
/2/ N.D.D'jačkovskij. Zvukovoj stroj jakutskogo jazyka. Č.1. Vokalizm.Jakutsk, 1971,pp. 120-124.
/3/ An oral communication by L.L.Gabyševa, who wrote a very interesting dissertation on this subject, however as yet unpublished.

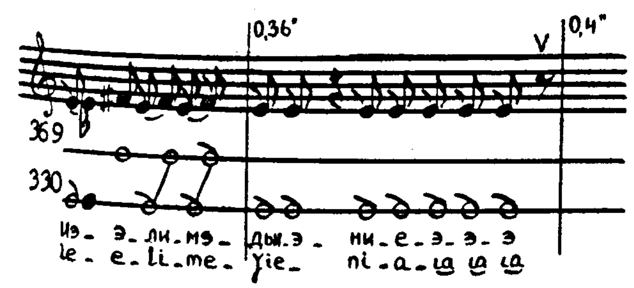
SUPPLEMENT

Fragments of the algys analysed (algys is a ritual request for a child's soul).
First, the record is given in the traditional music notation, then, in the new simplified notation, and finally goes the text. The position of intonemes next to the basic tones indicates fluctuations of the pitch zone.

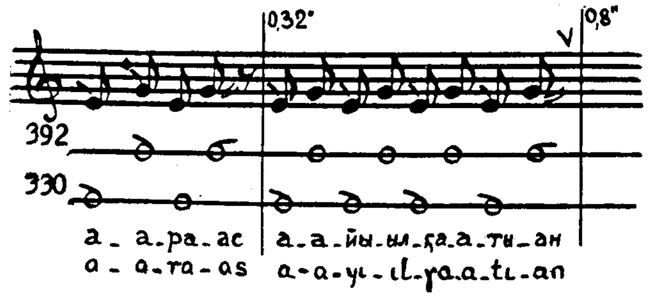
1. I line



2. 2 line



3. 7 line



Interaction between formant and harmonic peaks in vowel perception.

Hector Raul Javkin, Hynek Hermansky and Hisashi Wakita

Speech Technology Laboratory, Santa Barbara, California

ABSTRACT

The listener of a voiced vowel receives a signal consisting of formant-modulated harmonics. How this information is used in deriving both vowel timbre and resulting vowel identity is still not well understood. The suggestion by Klatt (1985,1986), that listeners perceive the actual resonance peaks, is contradicted by many works, including Mushnikov and Chistovich (1971) and Carlson, Granstrom and Fant (1975) who proposed weighted averages of neighboring harmonic peaks as the correlates of perceived vowel quality. Our perceptual experiments and re-analysis of the formant difference limen experiments of Flanagan (1955) and Nord and Sventelius (1979), support an interaction between formants and harmonic peaks in vowel perception.

INTRODUCTION

Although the influence of fundamental frequency on the perception of vowels is by now generally accepted [1,2,7,9, etc.], Klatt [5,6] has recently suggested that subjects respond to formant peaks without being affected by the location of the harmonic peaks determined by the fundamental, although he did find evidence for a normalization related to F0.

We found surprising support for the role of harmonic peaks in vowel perception in difference limen data shown in figures 1 [3], and 2 [8]. Both works provide the original measurement points along with the interpolated sensitivity curves. The measurements required a generous amount of interpolation to obtain smooth curves. If one takes into account the frequencies of the harmonic peaks in examining the published graphs, the origin of some of the outlying points can be hypothesised. The experiments were carried out with analog circuitry which might have produced some errors in fundamental frequency setting. If we allow for slight deviations of F0 values, almost all outlying points can be hypothetically attributed to the harmonic peak spacing. In Hermansky and Javkin [4] we reported on

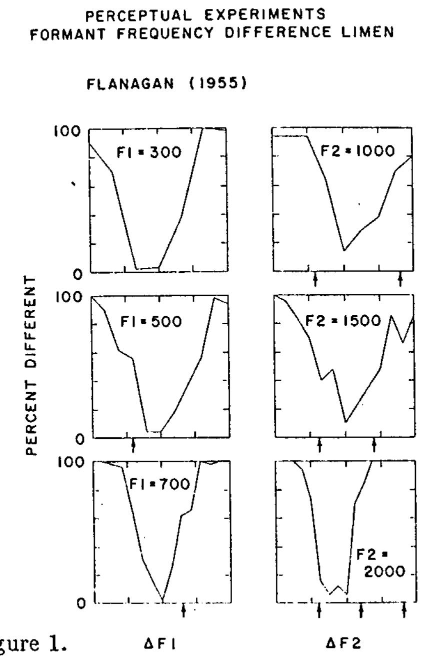


Figure 1.

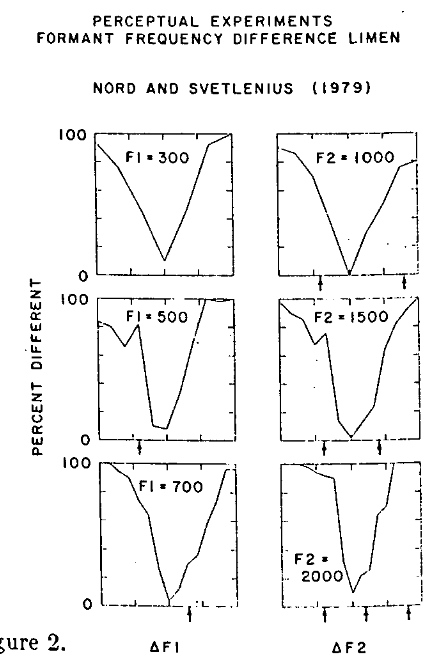


Figure 2.

further difference limen experiments on selected vowels that generally confirmed this hypothesis. Figure 3 shows the results for one of these experiments, for a vowel with formants at 500, 2000, 2500, 3500 and 4500 Hz. Bandwidths were 50, 90, 120, 150 and 180 for these formants. The fundamental period was varied between T0 = 8.5 and T = 8.0 msecs. in 0.1 msec. increments. Dashed lines in the figure connect points with equal formant frequency deviation from the reference vowel. Asymmetries, resulting from different distributions of harmonic peaks depending on the fundamental, are quite substantial. Figure 4 shows the results of the experiment with the same vowel but with the fundamental period T0 = 4.2 msec (with consequently wide harmonic spacing). Here the sensitivity curve shows an irregular (non-unique value) portion, similar to those observed in Flanagan's [3] data, coincident with a harmonic.

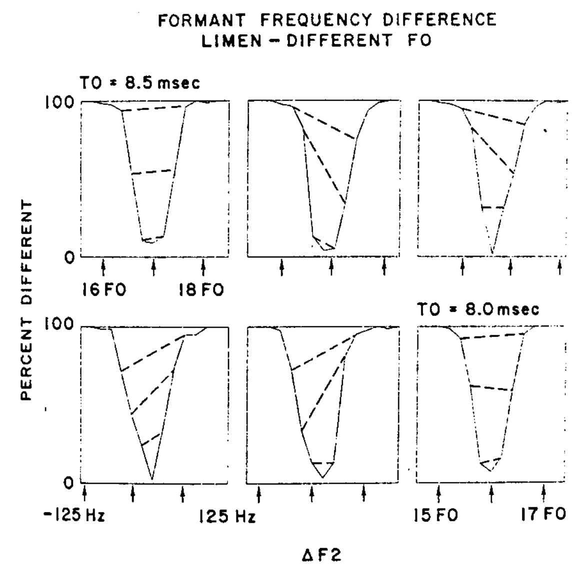


Figure 3.

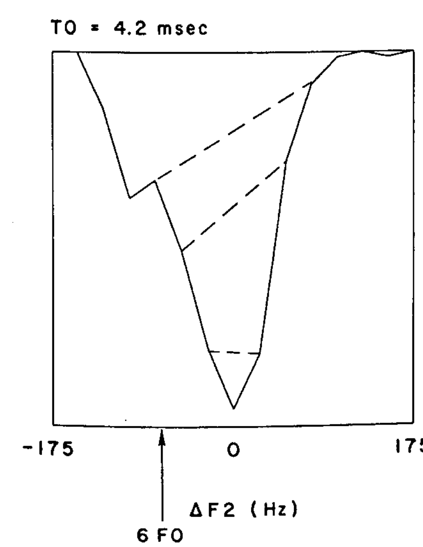


Figure 4.

The results provide further evidence for the hypothesis that the human auditory system tends to shift the formant peak estimate towards the nearest harmonic peak, but do not provide a basis for quantifying this shift. Carlson, Fant and Granstrom [1] attempted to model human listeners' perception of formant peaks and proposed the idea of "most important frequency" or MIF, which determines the weighted means of the two most prominent harmonics by an equation which can be written as follows:

$$\mathrm{MIF} = \frac{f_m W_m + f_n W_n}{W_m + W_n}$$

f_m is the frequency of the most prominent harmonic, f_n is the frequency of the next most prominent harmonic, and W_m and W_n are the weights given the respective harmonics. Carlson et al made the weights equal to the amplitude of the harmonics in the Sone space, so that S_m and S_n were used for W_m and W_n.

This formula suggests that listeners will accurately find the peaks when a formant lies between two harmonics, but will be less accurate when a formant coincides with or is close to a harmonic. An average of the two strongest partials, even one that is weighted towards the stronger, contains at least some contribution of the second strongest partial, and pulls the model's calculation of the formant away from the formant peak. If listeners can accurately find the formant peak when it coincides with a harmonic, the model will differ from their responses. The hypothesized estimation contains another, implicit hypothesis. Taking an arithmetic weighted mean in the sone space makes the assumption that listeners evaluate the amplitude of two neighboring harmonics for the purpose of peak location in the same way that they evaluate the amplitude of sounds presented separately. That is to say, they evaluate the relative amplitudes of the two without an interaction that increases the perceived amplitude of one or diminishes the perceived amplitude of the other.

Carlson et al conducted a perceptual experiment with F0 values from 100 to 160 Hz in 15 Hz steps and F1 values ranging from 250 to 350 Hz in 25 Hz steps. Their results showed that their hypothesis worked the best among those examined, although its prediction is quite different from the perceptual data when F0 = 100 and F1 = 300, i.e. when the formant coincided with one of the harmonics.

A less compressed scale such as magnitude or intensity will increase the contribution of the stronger partial and can increase the correlation between the output of Carlson et al's equation and their experimental data. It should be noted, however, that using a less compressed scale is functionally similar to using the same scale but with the addition of some form of peak enhancement.

MATCHING EXPERIMENT

To test whether a different scale would yield results closer to those of human listeners, a matching experiment was conducted. Our aim here was to avoid the effects of categorization that occur in vowel perception and investigate the psychoacoustic effects. Accordingly, single-formant stimuli with a single resonance driven by a pulse train with a flat spectrum were synthesized. F0 was kept constant at 200 Hz. One set of stimuli had peaks ranging from 600 to 800 Hz in 20 Hz increments, the other set had the same increments, but ranging from 2000 to 2200 Hz. Both sets were prepared with three bandwidths, of 50, 100, and 150 Hz, for a total of 66 stimuli. Durations of the single-formant stimuli were 500 msec with 40 msec leading and 70 msec trailing edges, while the tones had a 500 msec duration but 60 msec leading and 120 msec trailing edges. The inter-stimulus interval was 200 msec.

The presentation of the stimuli and the recording of responses were performed by a computer with a 16-bit digital-to-analog converter using a sample rate of 10 kHz with the output appropriately filtered. The stimuli were presented in different quasi-random orders to different subjects, who listened through earphones inside a sound-treated room. Subjects set the loudness of presentation to a comfortable level, and the level was checked visually after each subject completed the experiment. None of the subjects reported any hearing pathology. For each trial, subjects heard one of the single-formant stimuli followed by a sine wave. Their task was to match the timbre of the first stimulus by adjusting the frequency of the sine wave, using keys on a computer terminal. Their responses were limited to between 550 and 850 Hz for the F1 range stimuli and between 1950 and 2250 Hz for the F2 range stimuli. They could make adjustments for as long as they wished and heard a repetition of the two stimuli after each adjustment. When they indicated satisfaction with a match, their last adjusted value was automatically recorded in a computer file and the next trial began.

RESULTS OF MATCHING EXPERIMENT

Twelve subjects participated in the experiment. The task proved quite difficult for some subjects and two were eliminated after complaining of the difficulty and giving over a third of the responses at the response limits. The results for the different bandwidths did not differ significantly but were noisy. The results were band-limited to within 150 Hz (approximately two standard deviations) of the presented stimuli in order to limit somewhat the distorting effects of outliers. This meant that, for example, responses greater than 750 Hz to a 600 Hz stimuli were dropped from the data. Because of the band limitations in the presented stimuli and in the possible subject responses, points far away from the stimuli would severely distort the means. In addition, given a fundamental frequency of 200 Hz, a response of more than 750 Hz to a stimulus with a formant at 600 Hz might be the result of approaching the harmonic at 800 Hz. The results for the three bandwidths were combined in table 1, showing the results of the experiment for stimuli in the F1 range, and in table 2, showing the results for stimuli in the F2 range.

Table 1.

	Stimuli										
	600	620	640	660	680	700	720	740	760	780	800
subj	608	615	627	640	656	693	698	730	743	778	779
sone	640	648	661	674	687	700	713	726	739	752	760
mag	622	630	645	663	681	701	720	738	756	771	778
int	603	606	616	635	664	702	738	767	785	794	797

Table 2.

	Stimuli										
	2000	2020	2040	2060	2080	2100	2120	2140	2160	2180	2200
subj	2005	2010	2035	2047	2056	2118	2100	2140	2188	2165	2200
sone	2054	2059	2067	2077	2087	2098	2109	2120	2130	2138	2143
mag	2037	2043	2053	2067	2082	2098	2114	2128	2142	2153	2159
int	2010	2014	2024	2040	2065	2095	2126	2153	2172	2183	2188

Figures 5 and 6 graph the same results. Subjects' responses are represented by a solid line. The predictions of MIF calculated in sones are represented by a line of long dashes; the predictions calculated in magnitude are represented by a line of short dashes; and the predictions of MIF in the intensity space are shown by alternating short and long lines.

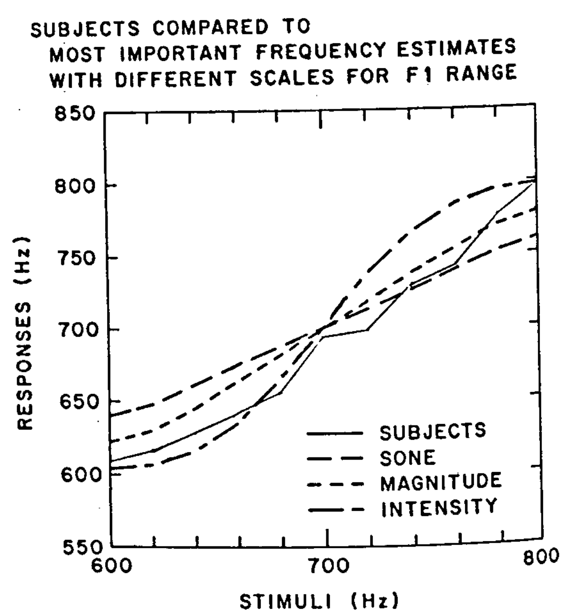


Figure 5.

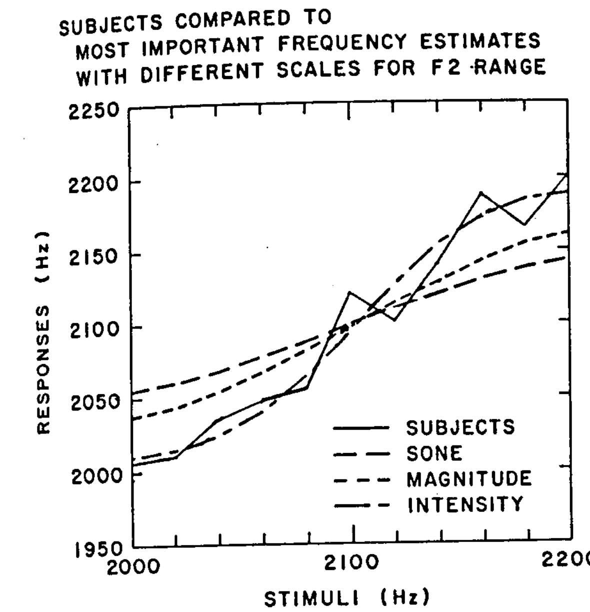


Figure 6.

The results for both sets are similar at their endpoints, although the data for the F2 range shows a less smooth pattern than the data for the F1 range. Both show a tendency for stimuli with harmonics close to the formant peak to attract responses and also for responses to show a "plateau" when the formant is equidistant between harmonics. The responses would approximate a straight line if subjects were responding to the location of the formant peak without regard to the location of harmonics, so that the experiment confirms the effect of harmonic peaks. Nevertheless the experiment does not confirm the predictions of Carlson et al.

CONCLUSIONS

It is clear from the experiments reported here, as well as the vast majority of the experimental literature, that the location of harmonics plays a role in the perception of vowels, and, more specifically, that harmonic peaks which coincide or nearly coincide with formants tend to attract judgments of formant location. This effect appears to be too strong to be represented by a weighted average of the two most prominent harmonics in the loudness space. Such an average can be improved by using a different scale, effectively expanding the differences in amplitudes. Although the results reported here are still somewhat sketchy and must be considered with caution, they support the idea that such an expansion is necessary to describe the response of the human auditory system.

REFERENCES

1. Carlson, R., Fant, G., Granstrom B. 1975. Two formant models, pitch and vowel perception - in Auditory Analysis and Perception of Speech (G. Fant & M.A.A. Tatham, eds.) Academic Press, London.
2. Chistovich, L.A. and Chernova, E.I. 1986. Identification of one- and two-formant steady-state vowels: a model and experiments. Speech Communication 5:3-16.
3. Flanagan, J.L. 1955. A difference limen for vowel formant frequency. Jour. Acoust. Soc. Am. 27.3:613-617.
4. Hermansky, H. and Javkin, H.J. 1986. Evaluation of ASR front-ends using synthetic speech - Paper presented at the 112th Meeting of the Acoustical Society of America, Anaheim, California.
5. Klatt, D. 1985. The perceptual reality of a formant frequency. Jour. Acoust. Soc. Am. 78, Suppl. 1:S81.
6. Klatt, D. 1986. Representation of the first formant in speech recognition and in models of the auditory periphery - Proceedings of the Montreal Symposium on Speech Recognition, McGill University, July, 1986.
7. Mushnikov, V.N., Chistovich, L.A. 1971. Method for the experimental investigation of the role of component loudnesses in the recognition of a vowel. Akusticheskii Zhurnal 17.3:405-411.
8. Nord, L., Sventelius, E. 1979. Analysis and perception of difference limen data for formant frequencies. STL-QSPR 3-4/1979:60-72.
9. Traunmuller, H. 1981. Perceptual dimension of openness in vowels. Jour. Acoust. Soc. Am., 69.5:1465-75.

ACKNOWLEDGEMENTS

We would like to thank Ted Applebaum, Jared Bernstein, Gregory De Haan, Brian Hanson, Dennis Klatt, Katia McClain, Lucio Mendes, Paul Neyrinck, Ben Reaves and Kathy Sangster for valuable discussions and help in various aspects of this paper.

THE STUDY OF AUDITORY DETECTION OF THE JUMP OF FORMANT FREQUENCY AND AMPLITUDE AS A CONSONANT

V.V. Lyublinskaya, E.I. Stolyarova, S.Y.Zhukov

Pavlov Institute of Physiology of the Academy of Sciences of the USSR, Leningrad, 195164

ABSTRACT

The perception of rapid changes (jumps) of formant frequency and amplitude in the spectrum of synthesized vowel was studied (in the experiments).

The boundaries of these changes associated with the consonants of different phonetic qualities were determined. Auditory images of studied stimuli in the form of space-time distribution of the responses of the detectors of amplitude irregularities in analyser frequency channels were received on the model. The character of the model representation of acoustic transitions from consonant to vowel was revealed.

INTRODUCTION

The subject of the study is auditory representation of "acoustic events" inherent in combinations of consonant and vowel phonemes in current speech. On dynamic spectrograms in these points one can observe rapid changes of formant frequency and amplitude, as well as those of the envelope amplitude.

It is known that the result of auditory analysis of formant transitions is used by man as phoneme determiners of diphones. As for automatic analysis of transitions it is known to be a difficult task, in particular, the formant frequency determination. For this reason it seems useful to apply the well-known principles of auditory processing for the analysis of transitions in speech signals.

According to some neurophysiological research the neurones in auditory system respond in a special way to the rapid changes in amplitude or spectral characteristics that occur in speech. The neurones which respond only to the positive or negative amplitude jumps (on- and off-responses) have been described in many papers/4/.

The simulations of such reactions was realized by the functional model of auditory determination of the amplitude irregularities (ADAI) /3,5/. It includes a model of peripheral spectral analyser (the "cochlea") and the system of the envelope processing in every frequency channel.

Positive and negative markers strictly localised in time are the responses to the respectively amplitude increase and decrease in the channels. The signal is represented in the ADAI model as the space-time distribution of the positive and negative markers in the channels. It was assumed that the markers might be used to form the segmentation function of speech flow and to sample the spectral information /3/. For this purpose, it is necessary to assume the integration of similar markers over the frequency channels. At the same time, on-and off-responses to narrow frequency signals may be strictly localized in frequency scale. This was also confirmed by the psychoacoustic data /1/. The narrow time and frequency localization of these reactions assumes the formation of space-time distributions as the response to the formant transitions.

This work was aimed to find out the possibility to use the responses of the ADAI model for the analysis of such acoustic events as the formant frequency and amplitude jumps. The present research has been inspired by the well-known fact that the jump of the formant frequency or the amplitude jump along the vowel-like segment of the signal is identified as a consonant and the whole signal as the syllable CV or VC depending on the direction of the jump /2,3/. The jump value determines the phoneme quality of the consonant. When the jumps are relatively large the stimulus is perceived as [m]V or [n]V, when the jumps are smaller - as [l]V /2/. The present research comprised 2 stages. Psychoacoustic experiments with synthesized vowels were carried out during the first stage. They were devised to determine the physical value of the jump of formant frequency and amplitude /$\triangle$F1 or$\triangle$A1/, when they were identified as the consonants [m] or [l] . The stimuli with the studied characteristics were analised in the functional model of peripheral spectral analyser and in ADAI model during the second stage.

PERCEPTION EXPERIMENTS.

Synthesized two-formant vowels (192 ms-24 pitch periods, 8 ms each) were used in experiments. The parallel formant analog synthesizer generated the stimuli. The variations of stimuli parameters were realized in 2 ways, as shown in figure 1. The parameters $F1_c$ and $A1_c$ of the first segment (the "consonant" segment, 64 ms) were controled. The second formant was constant and it was 10 dB less than the level of the first one. The set of experiments has been done, each test included only one type of stimuli. The values of F1 and F2 of the synthesized vowels are shown in the Table.

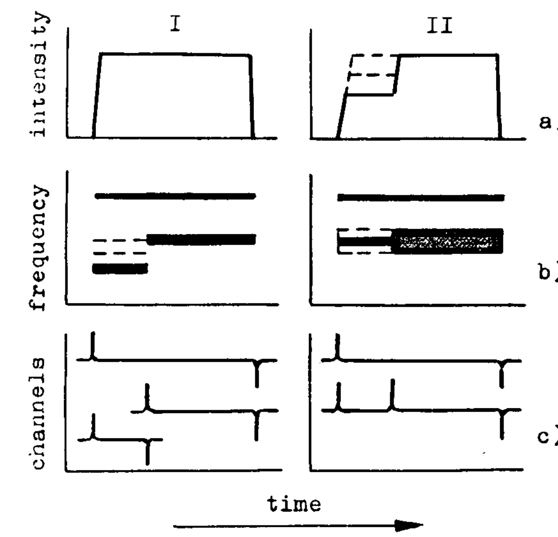


Fig.1. Structure of the stimuli in experiments: a) the amplitude envelope; b) the formant trackes; c) the markers in the channels of the ADAI model.

Table
The parameters of the stimuli and experimental data.

VOWEL	F2 (Hz)	$F1_V$ (Hz)	I ΔF(Hz) [m]	I ΔF(Hz) [l]	II ΔA(dB) [l]	II ΔA(dB) [m]	SUBJ.
u	625	400	120	60	4.4	11.1	1
			80	50	3.4	7.6	2
ɨ	1480	400	120	60	5.7	14.0	1
			100	60	4.8	11.2	2
i	2250	400	120	50	6.7	13.7	1
			100	60	4.3	12.9	2
e	1800	440	140	60	5.7	13.5	1
			130	50	4.3	11.9	2
o	780	535	175	55	4.3	9.6	1
			165	95	3.8	8.6	2
ε	1665	585	205	75	5.7	10.9	1
			205	115	3.8	11.5	2
a	1100	900	500	160	5.0	5.5	1
			500	210	3.8	9.0	2

Two methods - adjustment and identification - were applied in experiments. In the first case the subject controled the values of $F1_c$ and $A1_c$ to achieve the perception of the stimulus as [m] V or [l] V.

Results of the adjustment were registered by the experimenter.

According to the second method the sets of stimuli were presented to the subject, $\Delta F1$ and $\Delta A1$ being varied within definite limits.

Two subjects participated in adjustment experiments and five subjects - in identification experiments.

The results of the first type experiments are shown in the Table where the average values of $\Delta F1$ and $\Delta A1$ are indicated as the responses of each subject, when the stimuli were determined as the [l] V or [m] V syllables. The identification experiments data are analogous and therefore not described here.

The main properties of the perception of the jumps of formant frequency and amplitude are the following:

1. The perception of F1-jump depends on the quality of the vowel. The higher $F1_V$, the larger the jump $\Delta F1$, percepted as the consonant must be.
2. No regular dependence on $F1_V$ in the perception of the A1-jump is revealed.
3. The common feature inherent in perception of both frequency and amplitude jumps is revealed. The larger jump was identified as an [m], the smaller one as an [l] .

MODEL RESPONSES TO THE STIMULI

The sets of stimuli phonetically identified as V, [l] V, [m] V according to $\Delta F1$ and $\Delta A1$, were chosen for the analysis in the model. The markers distribution at the moment of the jump was examined for each stimulus. Two modes of operation were possible depending on the threshold value of markers generator: at the threshold of the detection of amplitude irregularity in the signal or at the threshold of the detection of the consonant while changing the amplitude of the signal.

Under the first condition the responses of the ADAI model were distributed on a wide frequency ranges. Under the second condition the markers were obtained in narrow frequency ranges near F1 and F2. We calculated and compared the number of channels were the markers could be registered at the moment of the jump. The number of marked channels correlate with the value of $\Delta F1$ or $\Delta A1$ under both conditions.

The patterns of the markers distributions were different for the frequency and amplitude jumps: only positive markers near $F1_V$ were registered for the amplitude jump, at the same time, the positive markers near $F1_V$, as well the negative ones near $F1_c$ were obtained for the frequency jump.

DISCUSSION AND CONCLUSION

The ADAI model reveals cues for the distinction between frequency and amplitude jumps, on the one hand, and allows to estimate the values of these both jumps according to the results of these experiments. We hope that the model features can help to describe the formant transitions in speech signals. The experimental data don't allow to make a conclusion about the information used by man for the phonetic interpretation of the frequency jumps. Whether he uses the time--frequency distribution of on- and off--responses only or he follows also the formant tracks. Possibly, both processings are necessary to provide the effective auditory perception of speech signals.

REFERENCES

I. Акустика речи и слуха: Сборник научных работ/Ред.- Л.А.Чистович./ - Л.:Наука, I986. - I44 с.
2. Жуков С.Я. Жукова М.Г. Восприятие слога в зависимости от его формантной структуры. - Физиология человека, I978, т.4, с.220-224.
3. Физиология речи. Восприятие речи человеком./ Чистович Л.А., Венцов А.В., Гранстрем М.П. и др./ Л., I976, 386 с. (Руководство по физиологии).
4. Delgutte B. Some correlates of phonetic distinctions at the level of the auditory nerve.- In: The representation of speech in the peripheral auditory system. Amsterdam, 1982, p.43-60.
5. Koshevnikov V.A., Stoljarova E.I. Segmentation of speech by a model of auditory system. - Symposium Franco--Sovietigue sur la parole, Lannion, 1983, p.95-105.

PERCEPTION OF FIRST AND SECOND FORMANT FREQUENCY TRAJECTORIES IN VOWELS*

Caroline B. Huang

Department of Electrical Engineering and Computer Science and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
USA

ABSTRACT

Previous studies suggest that the first formant trajectory in vowels is perceived differently from the second formant trajectory. F1 may be perceived as a weighted time-average of its time-varying frequency values (Huang, 1985, Di Benedetto, 1987). F2 in high vowels may be perceived with an overshoot (Lindblom and Studdert-Kennedy, 1967). The present study examines F2 in the low vowel region using synthesized utterances. Results from identification tests suggest that F2 in low vowels is perceived with an overshoot of 60 Hz in some contexts. However, results from preliminary experiments in which subjects matched vowels in nonsense words to steady state vowels seem to conflict with the perceptual overshoot theory for F2.

INTRODUCTION

The present study addresses the question: Is the first formant trajectory in a vowel perceived in a different manner from the second formant trajectory? Does a person listening to a vowel with time-varying formant frequencies use one strategy to determine a single value for the vowel's height, which is related to F1, and another strategy to determine the vowel's backness, which is related to F2? Evidence from perceptual tests suggests that the strategies for F1 and F2 perception are indeed different.

There are also theoretical reasons which suggest that F1 and F2 could be perceived differently. F1 and F2 correspond to independent phonological features, high-low and front-back, respectively. The phonological features high-low and front-back (and therefore F1 and F2) have independent articulatory correlates, tongue body height and tongue body backness. Tongue movements in running speech may result in different coarticulation effects for F1 and F2 trajectories. In the vowel spectrum, the spectral prominence corresponding to F1 may be widened or obscured by nasalization, which introduces a pole-zero pair to the spectrum (Stevens et al. [8]) or breathiness, which increases the amplitude of the fundamental harmonic (Bickley [1]). The F2 spectral prominence is not subjected to such effects. The different acoustic characteristics of F1 and F2 could be mirrored in their perception.

*Supported by grants from NINCDS (Nos. NS-04332 and NS-07040)

The properties of the peripheral auditory system form the basis of an alternative reasoning for the possibility that F1 and F2 are perceived differently. F1 and F2 occupy different frequency bands in the vowel spectrum. The peripheral auditory system processes low frequency and high frequency sounds differently, as shown by the differences in the shapes of the tuning curves for auditory nerves which respond most strongly to low frequency sounds when compared to those for auditory nerves responding most strongly to high frequency sounds. By this reasoning, it may be hypothesized not only that the F1 and F2 trajectories are perceived differently from each other, but that any formant trajectory is perceived differently depending on whether it is high or low in frequency.

PREVIOUS STUDIES

Studies by Huang [3], Di Benedetto [2] and Lindblom and Studdert-Kennedy [6] can be interpreted as evidence for F1 and F2 trajectories being perceived differently. Each study consisted of a series of tests in which subjects were presented with the synthesized vowels in nonsense words and asked to identify the synthesized vowel by making a forced choice between two vowels or two classes of vowels.

In Figure 1, the F1 trajectories for equivalent stimuli in Huang's study are shown. On the basis of identification data from five subjects, each of the stimuli would be called /ɪ/ half of the time and /ɛ/ half of the time. Results for the /ʊ,ʌ/ continuum (not shown) were very similar. The F1 target frequencies of the equivalent stimuli differ by up to about 20 Hz in both vowel continua. The stimulus with the longer onglide and offglide had to attain a higher F1 target value to be perceived to be equivalent to the stimulus with the shorter onglide and offglide. These results are consistent with a theory of perceptual averaging of F1. Subjects seem to perceive an effective F1 frequency which is between the maximum and minimum frequencies attained in the formant trajectory. Unfortunately, in this

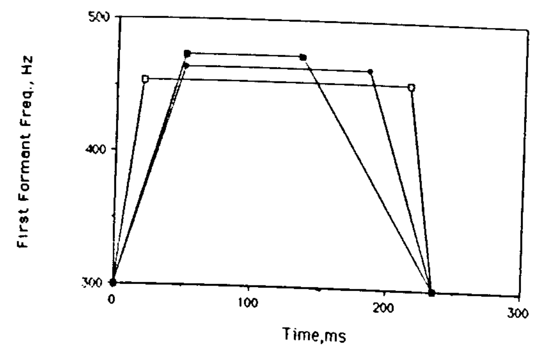


Figure 1: F1 trajectories for three equivalent stimuli in Huang's study.

study F2 was also varied, but only by half the change in F1 frequency on the Bark frequency scale (Schroeder et al. [7]). It may be argued that the change in F1 was perceptually more important.

In Figure 2, two vowel trajectories from Di Benedetto's study are shown. The F1 trajectory shape was different for two types of stimuli. The trajectories for F2 and all higher formants were the same and symmetric for both types of stimuli. Although the two F1 trajectories have the same average (defined as the area under the trajectory vs. time curve divided by the duration of the curve), they are perceived to be different vowels. The trajectory shape with the early steady state caused each of four subjects to identify the vowel as /ɛ/ more than half the time, and the trajectory shape with the later steady state was identified as /ɪ/ or /i/ more than half the time. The tendency was the same for three other subjects who were native speakers of languages other than American English, although the target value of the fifty-percent crossover stimulus was different. These results can be accounted for if a weighted average in which the early portion of the vowel is given more importance than the later portion is hypothesized. The later portion must be given non-zero weight, however, since it was shown in Huang's study that stimuli with trajectories as in Figure 1 with the same onglide duration and target frequency are not equivalent.

Lindblom and Studdert-Kennedy's study suggests that F2 is perceived with an overshoot. For example, for an F2 trajectory which rises to a target and falls again, subjects seem to hear an effective F2 frequency which is higher than the frequency actually attained. Note that if it is not hypothesized that F1 and F2 are perceived differently, Lindblom and Studdert-Kennedy's study would be seen to be in conflict with the studies described above. The vowel formants in their study had parabolic trajectories. The F1 trajectory was the same for all stimuli, while the F2 and F3 trajectories were either concave upward, resulting in a nonsense word of the form /jVj/ or concave downward, resulting in a nonsense word of the form /wVw/. The targets for F2 and F3 were varied while the target for F1 remained fixed for all stimuli, yielding a continuum between the vowels /ʊ/ and /ɪ/. Subjects' identification of the vowels with parabolic formant trajectories were compared to their identification of steady state vowels. The equivalent stimuli shown in Figure 3 are derived from the median fifty-percent crossover points in the identification curves for the steady-state vowels from ten subjects and median fifty-percent crossover shifts for the two contexts relative to the steady state vowels. (An identification curve shows percentage identification of a stimulus as /ɪ/, for example, versus the stimulus' position in the continuum.) The targets of the equivalent steady state and /wVw/ stimuli differ by 185 Hz. The targets of the equivalent steady state and /jVj/ stimuli differ by 75 Hz.

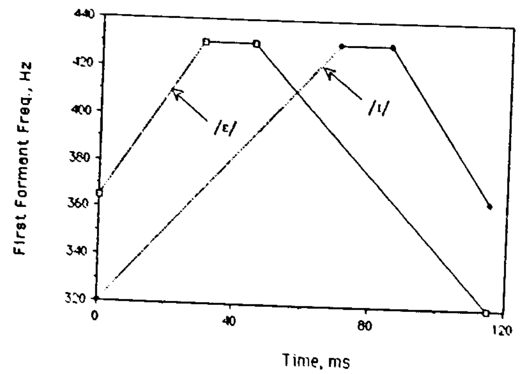


Figure 2: F1 trajectories from Di Benedetto's study. The vowels are perceived as ɛ and ɪ.

There was much inter- and intra-subject variation in Lindblom and Studdert-Kennedy's data which probably arose from subjects' difficulty in hearing the /wVw/ and /jVj/ stimuli as words. Huang [4] did a similar but smaller study using the nonsense words /əwVwə/ and obtained more consistent data which confirm Lindblom and Studdert-Kennedy's results.

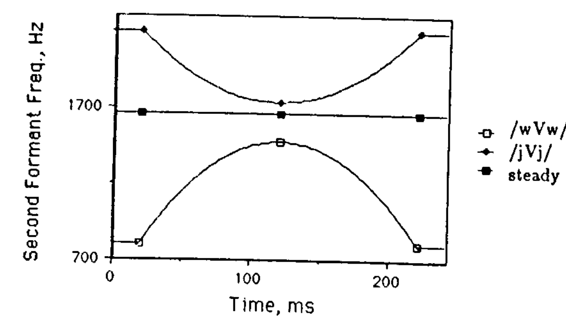


Figure 3: F2 trajectories for equivalent stimuli in Lindblom and Studdert-Kennedy's study.

F2 PERCEPTION IN THE LOW-VOWEL REGION

Lindblom and Studdert-Kennedy's study investigated F2 in the high-vowel region. The present study examines F2 in the low vowel region. Utterances of the form /əwVwə/ and /əjVjə/ were synthesized using the Klatt cascade formant synthesizer [5]. The target for the second formant of the vowel /V/ was varied in 57 Hz steps from 1090 Hz to 1720 Hz, a range of values appropriate for the vowel continuum /æ,ɑ/. The vowel had four formants, and the first, third, and fourth formant targets were 695 Hz, 2425 Hz, and 3500 Hz, respectively, for all stimuli. Two vowel durations were studied, 100 ms and 200 ms. The vowel trajectories in the nonsense words were parabolic and were concave upward for the /j/ context and concave downward for the /w/ context. Steady-state vowels with formant frequencies at the targets of the parabolic trajectories were also synthesized. The utterances were presented to five subjects in forced-choice identification tests in an order which ensured a balanced context. Each stimulus was repeated twelve times. Nonsense words of the same type and duration were presented together.

Fifty-percent crossover points were obtained by hand-fitting smooth curves to the identification curves. The fifty-percent crossover points for each subject and for the averaged identification curves are shown in Table 1. In Figure 4, the F2 trajectories of equivalent stimuli derived from the crossover points from the averaged data are shown. There is a shift in fifty-percent crossover point of 60 Hz

Subjects	Context /w/, 200ms /w/, 100ms	 steady, 200ms steady, 100ms	 /j/, 200ms /j/, 100ms
nd	5.2	5.5	6.8
	4.5	5.5	6.5
ms	6.8	6.5	8.2
	6.1	5.8	—
aw	5.8	7.0	7.5
	5.2	7.0	8.5
th	6.2	5.1	6.8
	5.5	4.6	8.1
cb	6.1	6.5	7.2
	5.5	6.8	7.2
Average	6.1	6.2	7.2
	5.5	5.8	7.6

Table 1: Results of the present study: 50% crossover points from identification curves. The numbers refer to the scale of stimulus numbers. Stimulus 1 was the most /ɑ/-like; Stimulus 12 was the most /æ/-like. The lower the crossover point, the more stimuli in the vowel continuum were called /æ/. The step-size was 57 Hz in F2. A dash (—) means data was unusable.

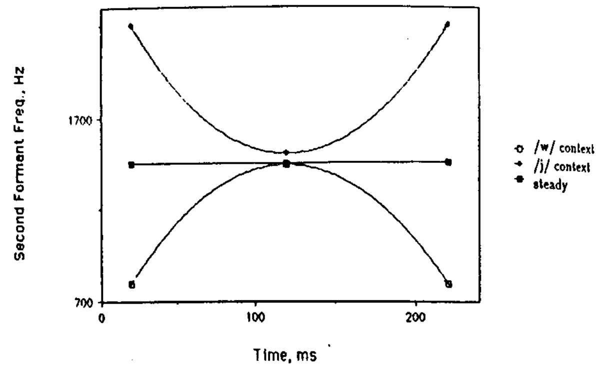


Figure 4: F2 trajectories for equivalent stimuli in the low vowel study.

when comparing the vowels in the /j/ context to the steady state vowels. The shift is in a direction consistent with a hypothesis of perceptual overshoot. On the average, there is no shift in the crossover point when comparing vowels in the /w/ context to the steady state vowels, since individual subjects showed shifts in both directions. There are small shifts in crossover point when comparing the 200 ms vowels to the 100 ms vowels in both the /w/ and /j/ contexts in directions indicating that the perceptual overshoot effect increases for shorter duration stimuli.

PRELIMINARY RESULTS FROM VOWEL MATCHING EXPERIMENTS

The same five subjects were then asked to match the vowels in the nonsense words to steady state vowels. F1, F3, and F4 of the steady state vowels for matching were at the target frequencies of those formants in the vowels in the nonsense words. The F2 of adjacent steady state vowels differed by about 30 Hz, and subjects knew the relative position of each matching stimulus on the steady state vowel continuum. Nonsense word stimuli were chosen in which subjects had unambiguously identified the vowels. The subjects matched the vowels by playing any desired vowels in sequence on a computer as often as they wished. As shown in Figure 5, subjects tended to match a vowel in the /w/ context to a steady state vowel with a lower F2 than actually attained in the parabolic trajectory, suggesting that F2 is averaged. Subjects also tended to match a vowel in the /j/ context to a steady state vowel whose target was lower than actually attained in the parabola, which is consistent with the original hypothesis of perceptual overshoot.

If F2 were perceived with averaging in the /w/ context, the vowel in /əwVwə/ should be equivalent to a steady state vowel whose F2 frequency is below that actually attained in the parabolic trajectory. That is, to be consistent with the trends seen in the preliminary vowel matching experiment, the fifty-percent crossover stimulus on the identification curve should be closest to the most extreme /ɑ/ stimulus for the steady state vowel continuum and closest to the most extreme /æ/ stimulus for the vowels in the /j/ context. The identification data for subjects TH and MS are consistent with the trends seen in the preliminary matching data. Identification data for the other subjects seem to conflict with this trend.

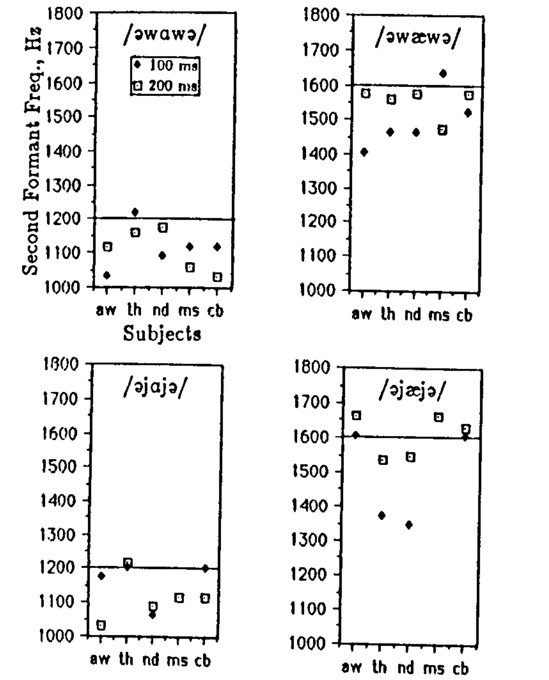


Figure 5: Data from the preliminary matching experiment. Horizontal lines show the F2 target value of the vowel in the nonsense word. Points show the F2 of the steady state vowel matched to the 100 ms and 200 ms parabolic vowels.

DISCUSSION

Apparent conflicts between the identification test results and the vowel matching results must be explained. The two kinds of experiments may be yielding information about different aspects of vowel perception. In this study, the matching experiment only investigated vowels which had been unambiguously identified by the subjects, while identification tests only yielded information about the vowels at the perceptual boundaries. A new vowel matching experiment must be done using the entire continuum of vowels in nonsense word contexts. The tasks of vowel identification and vowel matching are different, and this may also explain the apparent conflict. In vowel identification, a subject labels the vowel, possibly comparing it to an internal idea of how the vowel should sound. This "internal idea" may change depending on the context of the vowel. In vowel matching, a subject compares two "external" vowels and is not required to label. A subject may label the vowels before matching them, however. Subjects may listen to the vowels more analytically in the matching test than in the identification test, especially since they were allowed to play the vowels as often as they wished in this matching experiment. Subjects may use more language knowledge to perform the identification task than the vowel matching task. Individual subjects' strategies may account for the individual differences seen in the data.

Trying to determine whether the effects observed are a result of language learning or of properties of the peripheral auditory system is essential to understanding these effects. A starting point could be to see which of the observed effects can be reproduced using a model incorporating current knowledge of the peripheral auditory system.

Data from identification tests in previous studies and the present study are consistent with the hypothesis that the F1 and F2 trajectories are perceived differently. However, the original hypothesis that F1 is perceived with averaging and F2 with an overshoot does not account for all the effects observed in different types of experiments. Further work needs to be done to understand the relationship between the identification experiments and matching experiments for both F1 and F2 trajectories.

REFERENCES

[1] Bickley, C.A. Acoustic Analysis and Perception of Breathy Vowels. Working Papers, Speech Communication Group 1, Research Laboratory of Electronics, MIT, 1982.

[2] Di Benedetto, M.-G. An Acoustical and Perceptual Study on Vowel Height. Doctoral dissertation, Università degli Studi di Roma 'La Sapienza,' 1987.

[3] Huang, C.B., Perceptual Correlates of the Tense/Lax Distinction in General American English. Master's thesis, MIT, 1985.

[4] Huang, C.B. The Effect of Formant Trajectory and Spectral Shape on the Tense/Lax Distinction in American Vowels, *Proceedings ICASSP 86*, 1986.

[5] Klatt, D. Software for a Cascade/Parallel Formant Synthesizer. *JASA* 67(3):971-995, 1980.

[6] Lindblom B. and M. Studdert-Kennedy. On the Role of Formant Transitions in Vowel Recognition. *JASA* 42(4):830-843, 1967.

[7] Schroeder, M.R., Atal, B.S. and J. L. Hall. Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception. In B. Lindblom and S. Öhman (editors), *Frontiers of Speech Communication Research*, pages 217-229. Academic Press, London, 1979.

[8] Stevens, K.N., Fant, C.G.M. and S. Hawkins. Some Acoustical and Perceptual Correlates of Nasal Vowels. In R. Channon and L. Shockey (editors), *Ilse Lehiste Festschrift*, Foris Publications, Dordrecht, Holland, (in press).

ON VOWEL HEIGHT: ACOUSTIC AND PERCEPTUAL REPRESENTATION BY THE FUNDAMENTAL AND THE FIRST FORMANT FREQUENCY

Maria-Gabriella Di Benedetto (*)

Department of Information and Communication (INFO-COM)
University of Rome 'La Sapienza'- Faculty of Engineering
Via Eudossiana, 18- 00184 Rome, Italy.

ABSTRACT

Acoustic properties of vowels, which can be hypothesized to classify vowels along a dimension of height, are investigated. In particular, vowel representation in the (F1-F0) dimension (F1 and F0 are expressed in Bark) for five vowels of American English is presented and this analysis is compared with the analysis of the same speech materials in the traditional F1 vs F2 space. Results show that individual differences are reduced when the (F1-F0) dimension is used in the case of low vowels while for high and mid vowels the difference in F0 values among speakers is larger than that of F1 values. Perceptual experiments have been carried out using CVC and one-formant synthetic stimuli to examine the influence of F0 on the perception of vowel height. Results are in agreement with the observations on the acoustic analysis and suggest that either F1 and F0 are related in a more complex way than the (F1-F0) Bark-transformed difference or that the Bark scale should be modified at low frequencies.

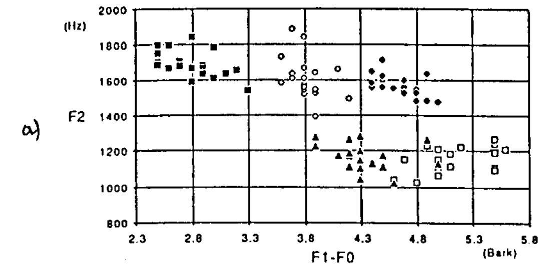
INTRODUCTION

Traditionally, vowel sounds have been classified along several dimensions: height, backness, tenseness, etc. The formant frequencies of vowels have been widely used as acoustic parameters representative of the different dimensions. For example, it is well known that the first formant frequency (F1) is an acoustic feature related to vowel height and the second formant frequency (F2) to vowel backness.
Syrdal (1985) has introduced the Bark-transformed (F1-F0) distance into a model for the auditory representation of vowels. Syrdal observes that the Bark-transformed (F1-F0) dimension corresponds to a dimension of vowel height. The results of Syrdal's analyses are in agreement with the perceptual results found by Traunmüller (1981). The latter proposes that the prevailing criterion for the perception of vowel height is the distance between F1 and F0 expressed in Bark, when F0 is not between 350 and 400 Hz, approximately.
The present study examines the effectiveness of the (F1-F0) distance to classify vowels according to vowel height. Acoustic analysis of five vowels of American English, in the (F1-F0) vs F2 space (F1 and F0 are expressed in Bark), is presented and compared with the analysis in the F1 vs F2 space. Perceptual experiments which have been carried out, using CVC and one-formant synthetic stimuli, to investigate the influence of F0 in the perception of vowel height are described. The agreement of the results obtained with the findings of the acoustic analysis and their interpretation are then discussed.

ACOUSTIC ANALYSIS

Experimental conditions and procedures
Five vowels of American English [I,ε,æ,a,ʌ] are the object of this analysis. In the vowel system of American English, these vowels are characterized by the feature (-round) and by being monophthongal, while the other vowels are all either (+round) or diphthongized. [I,ε,æ] are front vowels and [a,ʌ] are back vowels. [I] is (+high), [a,æ] are (+low), and [ε,ʌ] are (-high, -low). These vowels are considered in the context of voiced and voiceless stop consonants ([b,d,g,p,t,k]), forming CVC syllables, pronounced in the sentence frame "The ___ again". All the combinations between the vowels and the consonants listed are considered, except the non-symmetrical contexts with respect of voicing. In addition, hVd and #Vd syllables are analyzed. Three native speakers of American English, one female and two males, uttered the speech materials. They were asked to pronounce the sentences carefully and clearly. If a mistake occurs, the sentence is repeated. The sentences are pronounced in a random order. The set of syllables is pronounced three times. Thus, three versions of each vowel in each consonantal context are available. The speech materials are recorded in a sound-treated room using high quality equipment. The distance between the microphone and the speaker's mouth is about 20 cm. The recorded materials are then evaluated by a phonetically sophisticated listener. The speech signal is then stored on the MIT-Speech VAX-750. For this purpose, it is low-pass filtered at 4.8 kHz and sampled at 10 kHz.
The speech materials are analyzed using a software program KLSPEC developed by Dennis Klatt (1984). This program computes a 512-point DFT transform of slices of the signal (predifferenced and premultiplied by a Hamming window). The duration of the Hamming window is 30 ms at the sampling rate considered. In addition, fundamental frequency (F0) is determined by collecting frequencies of local maxima occuring below 3000 Hz and judging it to be that frequency (F0) which accounts for most peaks as harmonics. The program KLSPEC also calculates a spectrogram-like spectrum which is obtained by windowing a slice of signal (256 samples) and computing a 256-point DFT. A weighted sum of adjacent DFT sample energy is then computed for each of 128 spectrogram-like filters. Local maxima in this spectrum are most often indicative of the frequency positions of the formants. An interpolation algorithm improves the accuracy over the 40 Hz resolution implied by a 128-sample spectrum over 5 kHz. The spectrogram-like spectrum has been used for the estimation of the formant frequencies of the vowels under analysis. In some cases, in which this algorithm is not successful, the formant frequencies are manually extracted. DFT spectrum slices sampled every 5 ms are plotted and the frequency positions of the formants are evaluated by visual examination of the evolution of the locations of the DFT spectrum peaks in time. The temporal sampling point of F1, F2 and F0 is the time at which F1 reaches its maximum, as discussed in Di Benedetto (1987). The values of F0 and F1 are converted into a critical band tonality scale, according to Zwicker and Terhardt's (1980) mathematical approximation as adopted by Syrdal (1985).

Results of acoustic measurements
As expected, the highest F0 is found for the female speaker (CR) (191 Hz), while F0 for the two male speakers (JP) and (KS) is comparable (118 and 127 Hz, respectively).
The results of the analysis of the vowels [I,ε,æ,a,ʌ] for the three speakers considered in the (F1-F0) vs F2 space and in the F1 vs F2 space are extensively described in Di Benedetto (1987). In the present paper, results for only one of the speakers ((KS)) and one of the versions are presented as shown in Fig.1. Figure 1a shows that overlapping occurs in the (F1-F0) dimension only between [a] and [ʌ]. In the F1 vs F2 space (Fig.1b) overlapping occurs between [I] and [ε], [ε] and [æ], and [a] and [ʌ] while in the (F1- F0) vs F2 space, the [I], [ε] and [æ] areas are well separated. The use of the (F1-F0) dimension seems to improve the distinction between different vowels contiguous along the (F1-F0) dimension, for (KS). The results obtained for the other versions and speakers (Di Benedetto, 1987) show that similarly an improvement is obtained, in terms of better grouping and separation of the vowel areas in the (F1-F0) dimension, compared to what was obtained in the F1 dimension. However, problems of overlapping still occur between vowel areas of a single speaker in the (F1-F0) dimension. One should note that the differences in (F1-F0) values between vowels in voiced and voiceless consonantal contexts are lower than in F1 values (Di Benedetto, 1987). Consequently, one of the factors which contributes to a better separation of the vowel areas is that in the (F1-F0) dimension the vowel areas are better grouped.
The results of the comparison the vowel areas of the three speakers are summarized in Fig.2. Figure 2a (2b) shows for speaker (CR) (full losangues), (JP) (open losangues) and (KS) (full squares) the F1 and F2 values ((F1-F0) and F2 values) for each vowel, averaged over all the consonantal contexts and versions. The comparison of Fig.2a and Fig.2b shows that the difference in the representation of vowels for different speakers is reduced using the (F1-F0) parameter for the low and front vowel [æ] and the two back vowels [a,ʌ]. For the mid vowel [ε], in the (F1-F0) dimension the [ε]-area of the female speaker (CR) is shifted to lower values than those characterizing the [ε]-area of (KS) and (JP) and this effect is even more accentuated in the case of the vowel [I].
A comparison of these results with the analysis of American English vowels by Peterson (1961) has been carried out. It is noticed on Peterson's data that the difference in F1 values between male and female speakers, depends upon the range of F1 values. In particular, it is observed that this difference for high vowels is much smaller than for non-high vowels, and this difference increases when F1 increases. This result confirms what is observed in the present study.
Syrdal (1985) reports the Bark-difference means for ten vowels of American English on the Texas Instruments data base which consists of vowels in hVC words pronounced in isolation by 52 men, 51 women and 51 children. The data reported by Syrdal confirm that the difference in (F1-F0) values between male and female speakers depends on the height of the vowel considered. Note that both Peterson's and Syrdal's results are based on vowels pronounced in hVd or hVC words while the vowels of the present study are considered in several consonantal contexts.
In conclusion, acoustic analysis of the five vowels [I,ε,æ,a,ʌ] has shown that, in the dimension representing vowel height, individual differences for low vowels are reduced when the vowels are represented by the difference (F1-F0) rather than by F1. For high and mid vowels, on the other hand, a smaller shift in the F1 dimension would be needed to correct the differences in F0.



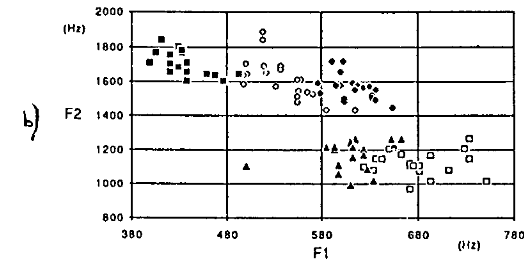
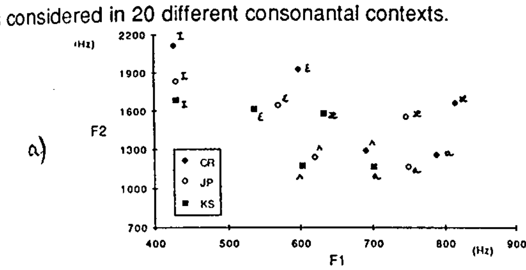


Figure 1: Results of the analysis in the a) (F1-F0) vs F2, and b) in the F1vs F2 spaces of the vowels [I,ε,æ,a,ʌ] (speaker (KS)). Each vowel is considered in 20 different consonantal contexts.



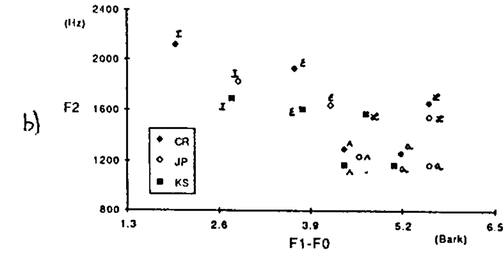


Figure 2: Average F1-F2 values (Fig.2a) and (F1-F0)-F2 values (Fig 2b) of the vowels [I,ε,æ,a,ʌ] for speakers (CR), (JP) and (KS). The averaged values are obtained by pooling all the consonantal contexts and versions.

PERCEPTUAL EXPERIMENTS

All the stimuli used in the experiment described were synthesized with the Klatt synthesizer (1980, 1984).

Experiment1
Description. The aim of this experiment is to investigate the influence of F0 on the perception of vowel height, using dVd synthetic syllables. One set of stimuli is characterized by F0=125 Hz (125-stimuli) while the two other sets of stimuli consists of stimuli which are identical to the previous ones as regards F1 and higher formant, while F0 of the stimuli of this experiment is increased in two steps: 60 Hz (185-stimuli) and 120 Hz (245-stimuli). Each set consists of 10 stimuli characterized by different values of F1 maximum, ranging from 300 Hz (stimulus #1) to 500 Hz (stimulus # 10) in steps of 30 Hz. Experiment 1 consists of two phases: a vowel identification test and a "boundary" identification test.
In the first phase the test was carried out on four american subjects. The subjects were all non-naive listeners, native speakers of American English and members of the Speech Communication Group at the Massachusetts Institute of Technology. They all name English as their best language.The stimuli used are the 125- and185-stimuli. The subjects were asked to identify the vowel of the synthetic utterances as [*i,e*] as justified by the results of a previous experiment (Di Benedetto, 1987).
In the second phase, 125-stimuli, 185-stimuli and 245-stimuli were used. Sequences of stimuli (and the same sequences in reverse order) characterized by the same F0 were played to the subjects who were asked to declare when their perception of the synthetic vowels presented changed from [*i*] to [*e*] or viceversa. Each sequence, in each order, was presented three times. Three subjects participated in this test.The subjects' description is identical to that of the subjects who participated in the vowel identification experiment
Results. Results of the identification test are presented for each subject, separately, in Fig.3 which shows that a change of 60 Hz in F0 does not result in a clear effect on the identification functions for any of the subjects who participated in the test. The three subjects who participated in the "boundary" identification test reported that they perceived the vowels of the synthetic utterances as either [*i*] or [*e*], as they were instructed. Figure 4 shows the results for each of the three subjects and indicates the first stimulus which is perceived as [*e*], when the sequences presented are ordered with ascending stimuli number, or the last stimulus which is perceived as [*e*], in the case of sequences ordered according to a descending stimulus number progression. Figure 4 shows that, in the case of the three subjects who participated in this test, an increase in F0 from 125 to 185 Hz does not result in a change of the perceptual boundary between [*i*] and [*e*], while a variation in F0 from 125 to 245 Hz does result in a consistent shift in this boundary. No difference was observed in the results obtained with sequences of stimuli with F1 increasing or in reverse order.

(*) this work was carried out while the author was with the Speech Communication Group at the Massachusetts Institute of Technology, Cambridge, MA, USA.

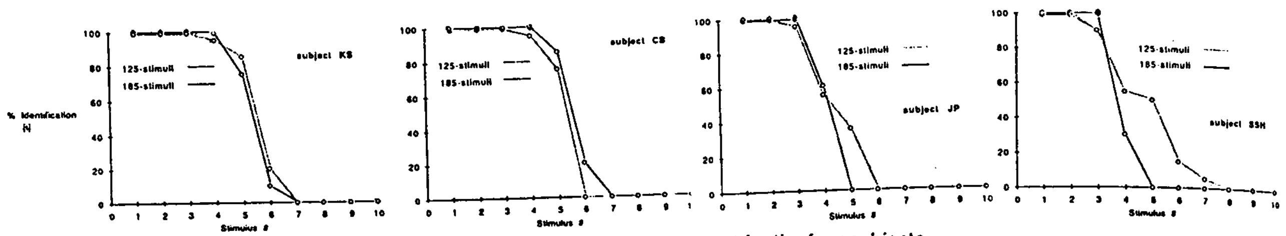


Figure 3: Results of the identification test for the four subjects.

Experiment 2

Description. The aim of this experiment is to investigate the influence of F0 in the perception of vowel height, using one-formant stimuli. Various one-formant stimuli with F0=125 Hz, 185 Hz or 245 Hz were generated. The one-formant stimuli with F0=125 Hz were characterized by five values of the formant (F1) (300, 350, 400, 500, 600 Hz). Each of these stimuli was matched against one-formant stimuli with F0=185 Hz and values of F1 ranging from the F1 value of the standard stimulus to the F1 value that would give the same F1-F0 for comparison and standard stimuli. Each pair was played three times. The same procedure was repeated with the same standard stimuli (F0=125 Hz) but the comparison stimuli were characterized by F0=245 Hz.Seven subjects participated in this experiment.They were non-naive listeners, native speakers of American English, and members of the Speech Communication Group at the Massachusetts Institute of Technology. They all named English as their best language. They were asked to indicate which pair of stimuli was most similar in terms of vowel height.

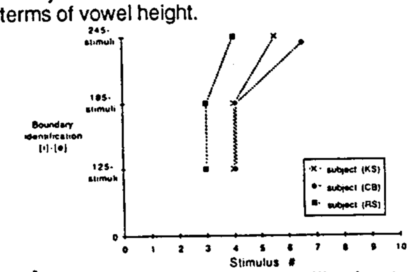


Figure 4: Results of the boundary identification test for the three subjects. Each dot on the figure (of different shape for each subject) indicates the stimulus at which the identification changes from [i] to [ɛ], in the case of the three stimuli F0 types.

Results. Figures 5 and 6 show the results obtained in experiment 2. Figure 5 shows on the abscissa the standard stimuli (with F0=125 Hz) identified by the F1 maximum value, and on the ordinate the comparison stimuli (with F0=185 Hz) which are matched against the standard stimuli. As shown on Fig.5 each standard stimulus is matched against three comparison stimuli: one with the same F1, one with the same (F1-F0) (in Hertz) and one with a F1 value intermediate between the same F1 and the same (F1-F0). For each standard stimulus, Fig.5 shows the value of F1 for best match in the case of each subject individually (1° column: subject (MA), 2° column: subject (TC), etc, as shown on the figure). A full (open) symbol indicates that the corresponding comparison stimulus was never (always) chosen as stimulus for best match by the subject. Partially open symbols indicate the percentage of times that this particular stimulus was chosen for best match. Figure 6 is similar to Fig.5 but indicates the results of the test in the case of the comparison stimuli with F0=245 Hz. In this case, each standard stimulus can be matched against five comparison stimuli: one with the same F1, one with the same (F1-F0) and three with intermediate values of F1, between the same F1 and the same (F1-F0). As in Fig.5, the value of F1 for best match is indicated by partial or complete blanking of the corresponding symbol, for each subject. Figure 5 shows that the F1 value for best match, in the case of stimuli with F0=185 Hz, corresponds to an exact formant match for low F1 values (300 and 350 Hz). For other values of F1 the match is in general between an exact formant match and values of F1 leading to similar (F1-F0) values. Note that in the case of the highest F1 value for the standard stimuli (600 Hz) the match is similar to (F1-F0) for subjects (TC) and (CH) and is close to this value for (KS). One should note that when F1 is high enough (for values higher than 400 Hz, approximately) F1 is out of linear Bark range. Consequently, the (F1-F0) distance expressed in Bark is always lower for comparison stimuli than for standard stimuli when F1 is in this range.

Figure 6 shows that the value for best match, in the case of stimuli with F0=245 Hz is in general at intermediate values of F1, between an exact formant match and values leading to similar (F1-F0) values for comparison and standard stimuli. In the case of the lowest values of F1 for standard stimuli (F1=300 Hz), the match is almost in all cases with stimuli characterized by F1=330 Hz corresponding to the first intermediate step. For values of F1 in the middle range (350, 400 and 500 Hz) the match shifts to stimuli with intermediate F1 values higher than in the case of standard stimuli with F1=300 Hz, with increasing F1 of the standard stimuli. Note, in fact, that for standard stimuli with F1=350 Hz the match is in general against comparison stimuli with F1=410 Hz and that for standard stimuli with F1=400 Hz or F1=500 Hz, the match is in general against comparison stimuli with F1=460~490 Hz and F1=560~590 Hz, respectively. The case of standard stimuli with F1=600 Hz is similar to the case of F1=400 Hz and F1=500 Hz, but note that for one subject (CH) the match is partially against stimuli with F1 values (720 Hz) leading to similar (F1-F0) values for comparison and standard stimuli.

DISCUSSION

Results of the perceptual experiments have shown that the influence of F0 in the perception of vowel height is related to F0 and F1-values. In particular, vowel identification experiments using CVC synthetic stimuli, have shown that an increase in F0 from 125 to 185 Hz does not result in a clear effect on the identification functions, while a variation from 125 to 245 Hz does result in consistently different judgements. A second experiment has been described, in which one-formant stimuli with F0=125 Hz and various values of F1 (300, 350, 400, 500, 600 Hz) were matched against one-formant stimuli in which F1 was adjustable and F0 equal to 185 or 245 Hz. Results show that the value of F1 for best match was usually between an exact formant match and a match yielding similar values of (F1-F0) for comparison and standard stimuli. The match was close to F1 for low F1 values and approached in general similar (F1-F0) values for higher F1. In some cases, in particular when comparison stimuli with F0=185 Hz were considered, the match reached the same (F1-F0) values (in Hertz) for comparison and standard stimuli. It has been noticed that in these cases, the (F1-F0) values expressed in Bark are lower for comparison stimuli than for standard stimuli.

The results of the perceptual experiments presented are in agreement with the observations on the acoustic analysis. The results of the acoustic analysis have shown that in the dimension representing vowel height, individual differences for low vowels are reduced when the vowels are represented by the (F1-F0) difference rather than by F1. For high vowels, the shift in the F1 dimension to account for differences in (F1-F0) increases the acoustic variability of the same vowel among speakers. For mid vowels, an intermediate effect is observed. In these cases (high and mid vowels), it has been observed that a smaller shift in the F1 dimension would be needed to correct the differences in F0. In the perceptual experiments, stimuli with three different values of F0 (125, 185 and 245 Hz) have been used. The average F0 value of the female speaker considered in the acoustic analysis is ~ 190 Hz and of the male speakers ~120-130 Hz, as previously mentioned. The results of the perceptual experiments for F0=125 Hz and F0=185 Hz, have shown that for low values of F1, F0 does not seem to influence the perception of vowel height. Correspondingly, one should note that it has been observed that the vowel area of the high vowel [I] for the male and the female speakers is located at similar values of F1. Experiment 2 has shown that when F1 is high, a change of F0 from 125 to 185 Hz influences the perception of vowel height and that stimuli with different values of F1 and F0 but similar (F1-F0) values are perceived as similar in terms of vowel height. Correspondingly, the acoustic analysis has indicated that the location of the [æ]-area corresponds to higher F1 values in the case of the female speaker, and to similar (F1-F0) values for the female and male speakers.

The interpretation of the results obtained can be given as follows. When F1 is sufficiently low (as in high vowels) and F0 assumes also low values (below ~200 Hz) F1 may be considered, by the perceptual mechanism which processes it, relative to the extreme end of the scale (the end of the scale is used as ananchor point) and is then the most relevant factor in vowel height perception. When F1 is high (as in low vowels) and F0 is sufficiently far from F1, F1 may be considered relative to F0 (not as previously to the end of the scale), F0 being used as an anchor point, and the distance between F1 and F0 (in Bark) is determinant in the perception of vowel height. When F1 is at intermediate values, or the distance between F1 and F0 is not large enough, F1 and F0 would both intervene in the perceptual process determining vowel height in a relation which would not attribute the same weight to F1 and F0. This interpretation would imply a non-uniform vowel normalization in agreement with Fant's study (1975).

This hypothesis finds support in results of physiological experiments carried out by Delgutte and Kiang (1984), as pointed out by Stevens (1985). These investigators have observed the location of the largest components in the discrete Fourier transforms of period histograms obtained from auditory-nerve fibers with various values of the characteristic frequency (CF). The stimuli were steady-state two formant stimuli with F0=125 Hz. Delgutte and Kiang note that for all vowels, there is a CF region which is located around F1 (F1 region) where the harmonics close to F1 dominate the response spectra. In addition, they observe that this region is flanked on the low-CF by another region in which the harmonics close to CF are the largest components in the response spectra. These harmonics correspond to the fundamental frequency or to intermediate values between F1 and F0. For low vowels, this region extends up to about 400 Hz while on the contrary, for high vowels, this region is not distinct. Delgutte and Kiang observe that "...the open-close dimension of phonetics correlates with both the position of the F1 region along the CF dimension and with the extent of the low-CF region". This observation could justify the results of the present study that for low F0 values, F1 determines the perception of vowel height when F1 is low (high vowels), whereas if F1 is high (low vowels) F0 influences vowel height perception. Unfortunately, Delgutte and Kiang do not present results in the case of higher values of F0. Consequently, the results of the present study in the case of higher values of F0 cannot be interpreted on the same basis. We want to point out that the perception of vowels with F1 and F2 closer than 3.5 Bark could be based on one equivalent formant located in a position intermediate between the two formants, according to the categorical perceptual effect SCG (Spectral Center of Gravity) found by Chistovich et al. (1979). It could be hypothesized then that this one formant is relevant, in the cases of vowels with F2-F1 < 3.5 Bark, to vowel height perception.

This aspect of the problem is not addressed in the present study. We want to suggest that our interpretation of the relation between F1 and F0 in the perception of vowel height is appropriate in the case of front vowels, but that for back vowels additional factors could be relevant, such as, according to the SCG theory, the relative amplitudes of F1 and F2.

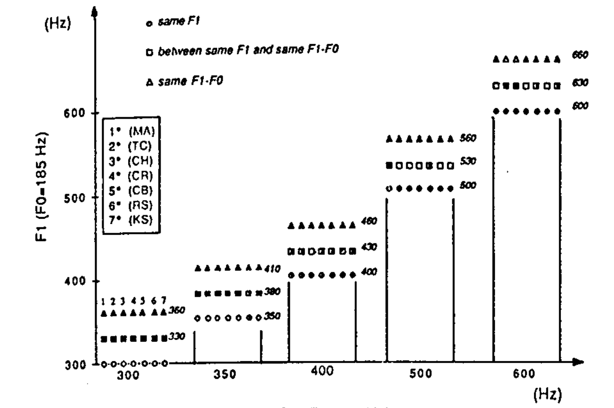


Figure 5: Results of experiment 2 for comparison stimuli with F0=185 Hz.

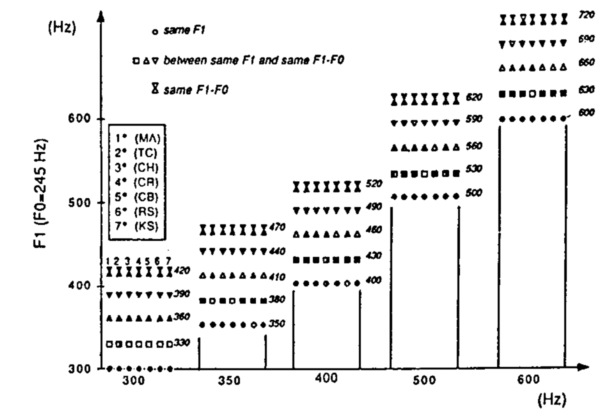


Figure 6: Results of experiment 2 for comparison stimuli with F0=245 Hz.

REFERENCES

Chistovich, L.A., Sheikin, R.L. & Lublinskaya, V.V. (1979) "Centres of gravity and spectral peaks as the determinants of vowel quality", in: B.Lindblom and S.Öhman, eds., *Frontiers of Speech Communication Research*, Academic Press, London, 143-157.

Delgutte, B. & Kiang, N.Y.S. (1984) "Speech coding in the auditory nerve: I. vowel-like sounds", *J. Acoust. Soc. Am.* 75 no.3, 866-878.

Di Benedetto, M.G. (1987) "An acoustical and perceptual study on vowel height", Ph.D. thesis, University of Rome, Italy.

Fant, C.G.M. (1975) "Non-uniform vowel normalization", *STL-QPSR* 2-3.

Klatt, D.H. (1980) "Software for cascade/parallel formant synthesizer", *J. Acoust. Soc. Am.* 67 no.3, 971-995.

Klatt, D.H. (1984) "M.I.T. SpeechVAX user's guide", preliminary version.

Peterson, G.E. (1961) "Parameters of vowel quality", *J. Speech and Hearing Res.* 4, 10-29.

Stevens, K.N. (1985) Personal communication.

Syrdal, A.K. (1985) "Aspects of a model of the auditory representation of American English vowels", *Speech Communication* 4, 121-135.

Traunmüller, H. (1981) "Perceptual dimension of openness in vowels", *J. Acoust. Soc. Am.* 69 no.5, 1465-1475.

Zwicker, E. & Terhardt, E. (1980) "Analytical expressions for critical band rate and critical bandwidth as a function of frequency", *J. Acoust. Soc. Am.* 68, 1523-1525.

PERCEPTION DE VOYELLES EN CONTEXTE NASAL DANS L'ESPAGNOL PARLE A PORTO-RICO

CARLOS A. CASABLANCA

Département des Humanités
Université de Porto Rico à Mayagüez
Mayagüez, Porto Rico 00709

ABSTRACT

Lors d'un test perceptif inspiré de la méthode "tape splicing", cent soixante-dix sujets Porto-Ricains se sont prononcés sur la qualité orale ou nasale de la voyelle de la derniere syllabe des mots dont la con sonne finale avait été effacée électronique ment. Les mots enregistrés étaient du genre VCVC ou VCVN et ceux proposés après découpa ge étaient du genre VCV. Chaque mot compor- tait au moins une paire minimale.
Les mots ont été proposés en deux séries dif- férentes, chaque série se différenciant u- niquement par la quantité de transition re- tranchée en même temps que la consonne. **Dans la premiere série seul 2,5 centièmes de sec. ont été retranchés**, tandis que dans la deuxième c'est la transition toute enti- ère qui a été éliminée.
Le rôle de la transition dans la perception a ainsi été mis en évidence et on a pu con- firmer la portée de l'action assimilatrice.

INTRODUCTION

De nombreux auteurs se sont intéressés à l' aspect perceptif de la nasalisation. Des chercheurs tels que Lintz et Sherman /1/, Malécot A. et G. Metz /2/ et Ali Latif /3/, comptent parmi ceux qui ont publié le ré- sultat de test perceptifs effectués sur des sujets français ou américains.
Mais les techniques mises en application, ainsi que les formats des corpus utilisés dans l'élaboration de leurs expériences dif fèrent grandement selon la langue étudiée, l'époque à laquelle l'expérience a été con- duite et le but recherché. Les résultats ob tenus ont pu, eux aussi varier en conséquen ce.
Ainsi Lintz et Sherman ont conclu que le de gré de nasalisation d'une voyelle en contex te nasal varie d'après la hauteur de la voy elle, les voyelles hautes étant les moins nasalisées et d'après le point d'articula- tion, les voyelles postérieures étant les moins nasalisées.
Les résultats du test conduit par Ali indi- quent que les consonnes qui suivent la voy- elle basse /a/ ont été perçues comme nasa- les plus souvent que les consonnes se trou- vant après /u/ et /i/. Dans le même article Ali démontre que dans le cas des syllabes du type CVVN le sujet peut détecter la pre- sence de la nasalité même lorsque la conson ne nasale ainsi que la transition ont été retranchées. Ceci est dû selon Ali à l'aper ture prématurée du chénal vélopharyngal an- ticipant la consonne nasale finale.
Dans la même perspective nous avons réalisé un test qui a pour but de mettre en éviden- ce dans l'espagnol de Porto-Rico les voyel- les que les consonnes nasales adjacentes af- fectent le plus.
Ce test perceptif s'inspire dans le domaine de la technique de la méthode dite "tape s- plicing" ou découpage électronique d'une bande enregistrée.
La théorie sousjacente qui permet ce genre d'expérince se fonde sur des études ciné- radiographiques qui montrent qu'un seul mou vement articulatoire peut correspondre à plus d'un segment phonétique.
Le concept de co-articulation à la base de notre travail a amené V.A. Kozhevnikov et L.A. Chistovitch /4/ à proposer que la syl- labe soit l'unité minimale d'articulation. La fusion intime entre les différents élé- ments d'une même syllabe se traduit presque obligatoirement par des phénomènes d'articu lation qui aboutissent à la création des va- riations allophoniques.
Dans le cas des segments ou des suites qui comportent un phonème nasal, il a toujours été admis que la coloration particulière de la consonne est perceptible déjà lors de l' énonciation de la voyelle orale. Moll et Daniloff /5/ ont démontré que dans des sé- quences du genre CVN le voile du palais est déjà abaissé dès le début de la voyelle o- rale.
Il importe d'établir maintenant et c'est le but du présent travail, si le partage du trait nasal est aussi courant dans l'espa- gnol de Porto-Rico que dans les autres lan- gues pour lesquelles nous avons déjà des résultats.

CORPUS UTILISE DANS LE TEST DE PERCEPTION

Les mots proposés dans le test de percep- tion sont tous bisyllabiques et ont été pré sentés en deux séries. Ils portent tous l' accent sur la dernière syllabe et ont été retenus car ils ont tous au moins deux pai- res minimales dont une se termine par une consonne nasale.

1ere série	2eme série
irán	irán
mantén	mantén
matón	matón
cojín	cojín
guión	guión
ladrón	
limón	
Corán	
harem	
caí	
ladró	
mantel	
irás	
Balí	
coral	

Aux mots de la première série, sous contrô- le de l'oscillographe, nous avons retranché toute la consonne nasale ainsi que 2,5 cen- tièmes de sec. de la transition.
Aux mots se terminant par consonne orale nous avons retranché le même segment qu'aux mots se terminant par consonne nasale. Aux mots se terminant par voyelle deux ou trois vibrations leur ont été retranchées; juste assez pour donner l'impression que "quelque chose" avait été enlevé.
Les mots de la deuxième série, se sont vu retrancher outre la consonne finale, toute la transition. Le but de cette opération é- tait de cerner le rôle de la transition.
Il convenait de savoir si les sujets porto- ricainsperçoivent le trait nasal lorsque toute la transition et une partie de la voy elle ont été enlevés.
Pour nous assurer que rien ne restait de la transition, dans certains cas nous avons dû couper jusqu'à trente pour cent de la voyel le.
Nous aurions voulu préparer le corpus à l' aide des cinq voyelles de l'espagnol, mais malheureusement les lois distributionnelles de l'espagnol empêchent de travailler con- venablement avec la voyelle postérieure /u/ qui est très rare en fin de mot.

Les sujets

Le corpus a été prononcé à vitesse normale par un sujet féminin, choisi parmi douze candidats pour sa diction claire et pour son accent typiquement porto-ricain. Cette précaution garantissait qu'il n'y aurait pas d'interférence linguistique à cause de l'anglais. Avant d'enregistrer le corpus la fréquence fondamentale de la voix du sujet a été calculée à l'aide d'un spectrogramme. Elle a été enregistrée à 140Hz.
Les sujets linguistiquement naïfs qui ont eu à se prononcer sur le corpus enregistré étaient au nombre de cent quatre-vingt-cinq. Pour diverses raisons, les formulaires de réponse de sept d'entre eux ont dû être an- nulés. Un nombre élevé de réponses garantit par la suite une analyse plus précise des résultats.
Les sujets devaient identifier la nature de la dernière consonne d'un mot préalablement enregistré. Des syllabes du type CVC et CVVC ont été utilisées, leur dernière con- sonne pouvant être /m/, /n/ ou un phoneme non nasal. La dernière consonne a donc été retranchée et les syllabes CV et CVV résul- tantes ont été proposées aux sujets qui de- vaient identifier la nature de la consonne manquante.

Résultats fournis par la premiere série.

Voici un tableau récapitulatif présentant les mots terminés par une nasale, qui ont été proposés aux étudiants, ainsi que le nombre de fois où ceux-ci ont identifié cor- rectement la nature de la consonne manquante.

matón	99/177	= 56,5%
guión	117/177	= 66,0%
ladrón	99/177	= 56,0%
limón	150/177	= 85,0%
mantén	48/177	= 27,0%
harem	69/177	= 39,0%
irán	81/177	= 46,0%
Corán	27/177	= 15,0%
cojín	66/177	= 38,0%

La nasalisation dans le cas de "limón", 'ci- tron', est évidemment élevée à cause du pho- nème bilabial situé devant la voyelle. La voyelle postérieure /o/ ne semble pas of- frir une résistance importante à l'assimila- tion nasale. La voyelle antérieure /e/ sem- ble avoir mieux résisté à l'assimilation que le /o/.
Nous nous attendions à ce qu'un plus grand nombre d'individus jugeât que le /a/ était fortement nasalisé; il n'en fut rien lors de cette première série de mots.
Aucun des sujets n'a pu identifier comme telles la totalité des consonnes nasales retranchées. Voici les résultats.

0	sujets ont identifié	9	nasales sur	9
15	"	8	"	9
21	"	7	"	9
27	"	6	"	9
69	"	5	"	9
24	"	4	"	9
15	"	3	"	9
3	"	2	"	9
3	"	1	"	9
0	"	0	"	9

Il est possible d'affiner les résultats. Pour ce qui est des mots auxquels seul une partie de la transition a été enlevée, 135 personnes sur 177 soit 76% des sujets ont percu le trait nasal au moins cinq fois sur neuf. Ces chiffres tendent à confirmer que lorsque la transition est présente, la nasalité est perçue avec une relative faci- lité même si la consonne nasale est oblité- rée. Si l'on compare ces chiffres à ceux

qu'avance Ali (op.cit) dans son tableau numéro deux page 539 nous constatons qu'il explique que tous ses sujets, vingt deux en tout, ont identifié correctement cinq nasales ou plus sur neuf. Dans ses résultats la transition a été complétement effacée.
Les résultats de Clumeck /6/ pour le français révelent que les sujets français n'ont pu déterminer avec certitude si les mots auxquels on avait retranché la dernière consonne étaient ou non terminés par nasale.
Les conclusions de ces deux expériences vont dans le sens de la théorie selon laquel le le degré de nasalité atteint par une voyelle orale diffère selon la langue. Il semble que nos résultats se situent entre ceux d'Ali et ceux de Clumeck.

<u>Résultats fournis par la deuxieme série.</u>

La deuxième série comporte cinq mots se terminant par une consonne nasale. Ces mêmes mots ont été utilisés lors de la première série proposée. La seule différence est que cette fois ci, toute la transition a été enlevée. Nous espérions ainsi dégager l'importance de la transition dans la perception du trait nasal. Voici les mots proposés et le pourcentage de réponses exactes.

matón	36/177	= 20%
guión	72/177	= 41%
mantén	39/177	= 21%
cojín	51/177	= 30%
irán	48/177	= 28%

Le pourcentage de réussite dans ce cas est moindre que lorsque la syllabe a été seulement privée d'une partie de la transition; 28% pour la deuxième série contre 47% pour la première série.
La plupart des sujets ont perçu les voyelles sous étude comme étant complétement orales. Les résultats statistiques montrent néanmoins que les sujets ont eu tendance à répondre au hasard. Avec la transition nous avons dû éliminer, dans la majorité des cas plus de 30% de la voyelle.
Remarquons que c'est pour le /o/ que les résultats diffèrent le plus. Lorsque toute la transition a été enlevée, la nasalisation a été remarquée environ 30% de fois contre 70% de fois lors de la série précédente.

CONCLUSION

Ce test perceptif laisse entrevoir que les auditeurs portoricains perçoivent dans les voyelles orales une coloration semblable au trait nasal qui normalement n'accompagne que les phonèmes nasaux. Cette coloration est importante surtout lorsque la voyelle orale se trouve dans une syllabe terminée par une consonne nasale, position dans laquelle elle subit toujours une assimilation regressive.
Tous nos sujets ont en effet perçu avec plus ou moins de netteté la teinte nasale surtout lorsque la transition n'a pas été retranchée dans sa totalité. L'élimination de toute la transition, ainsi que le raccourcissement de la voyelle qui en découle ont non seulement rendu plus difficile la perception de l'assimilation nasale, mais encore ont incité les sujets à répondre au hasard. Le rôle de la transition n'a donc pas pu être entièrement cerné.
Ces résultats montrent néanmoins que la voyelle /o/ située en contexte nasal est celle à laquelle on a le plus souvent attribué le partage du trait nasal.

BIBLIOGRAPHIE

/1/ Lintz et Sherman, "Phonetic elements and perception of nasality", J. Speech Hearing Research 4, 1961.
/2/ Malécot A. et G. Metz, "Progressive nasal assimilation in French", Phonetica 26, 1971.
/3/ Ali Latif et al., "Perception of coarticulated nasality", J.A.S.A. Vol.49 No2 Part 2, 1971.
/4/ V.A. Kozhevnikov and L.A. Chistovitch, cité dans J.A.S.A. vol. 50 No 2 part 2, 1971. (il s'agit de la construction CV)
/5/ Kenneth L. Moll et R. Daniloff, "Investigation of the timing of velar mouvements during speech", J.A.S.A. Vol. No 2 part 2 1971.
/6/ Clumeck H., "Degrees of nasal coarticulation", Monthly Internal Memorandum, Phonology Lab. University of California, Berkely, July 1971.

AN EXPERIMENT ON THE CUES TO THE IDENTIFICATION OF FRICATIVES

HARTMUT TRAUNMÜLLER DIANA KRULL

Institutionen för lingvistik
Stockholms Universitet
S-106 91 Stockholm

ABSTRACT

Synthetic fricatives with two spectral peaks scanning a wide range of frequencies were put into three versions of the context <a_'ɛ:>, also generated synthetically, and imitating a male speaker (1), a child (2), and an aroused male speaker (3) with elevated F_0 and F_1. The stimuli were presented in two orders, with increasing or decreasing frequencies of the spectral peaks, to 16 speakers of Swedish who identified the fricatives as <f>, <s>, <ɕ>, <ʂ>, or <ɧ>. In a given context, the obtained phonetic boundaries followed mainly the spectral peak lowest in frequency, while the upper peak contributed only marginally even if it was at a distance less than the "critical distance" of about 3 Bark. In context (2), as compared with (1), the phonetic boundaries were shifted up, but less (in Bark) than the vowel formants.

INTRODUCTION

It is well known that the characteristic frequencies, i. e., the frequencies of the formants and the fundamental in speech sounds with a given phonetic quality vary with the overall dimensions of the speaker's vocal tract. If the characteristic frequencies of vowels are converted into a measure of tonotopical place, such as critical band rate (Bark), differences in speaker size can be seen to correspond to a tonotopic translation of the auditory pattern of excitation <11>.

Identifications of synthetic two-formant vowels revealed that a uniform tonotopic compression of the auditory pattern of excitation with a fixed point in the region of F3 also preserves phonetic quality <12>. Natural vowels are transformed in this way in shouting and in whispering <11>.

The present investigation is about the transformations the spectra of voiceless fricatives can be subjected to without affecting their phonetic quality. It is known that voiceless fricatives can be synthesized satisfactorily with two resonances and one antiresonance and that the cues to the phonetic identity of voiceless sibilants reside mainly in the stationary part of their spectrum, while the transitions are more important for nonsibilants <5, 7>. One-parameter sibilants can be synthesized using a resonance and an antiresonance one octave lower in frequency <5>. Such sibilants lack intrinsic cues to speaker size. In spectrogram reading, the Swedish voiceless sibilants can be distinguished by the frequency of spectral energy onset while there is more variation, even within the same speaker and context, in the detail above that frequency <6>. A second characteristic spectral peak can, however, often be discerned and one question we address here is whether this second peak is used to normalize for speaker size. We also investigate in how far a vocalic context can serve this purpose.

METHODS

Subjects

The experiments were conducted with a group of 20 native and 6 non-native speakers of Swedish, all employees or students at the Institute of Linguistics at Stockholm University. None of them reported auditory handicaps and all were familiar with the phonetics of Swedish, possessing /f/, /s/, /ɕ/, and /ʃ/. We report here the results of 16 native speakers with uniform behavior, mostly speakers of the local variety with the distributional allophones <ʂ> and <ɧ> for /ʃ/, but including three speakers of southern varieties, who had no <s> in their own speech.

Stimuli

The stimuli were synthetic VCV sequences. The vocalic segments had been obtained by synthetic imitation of a natural <a's:ɛ:>, produced by a male speaker of Swedish (Stockholm variety). A three parameter voice source <3> signal in accordance with that utterance was generated by the procedure described in <12>. The vocalic as well as the fricative segments were generated in serial synthesis by use of a block diagram simulating program (sampling at 16 kHz, 16 bit/sample). Eight vowel formants were used. Their bandwidths obeyed the standard relation $B_i = 0.05\ F_i + 50$ Hz.

The fricatives were generated by feeding white noise through a high-pass and a low-pass resonance filter, both of second order and with Q=10. The two resonance frequencies F_l and F_h were varied in steps of a factor $4^{1/9}$ (approx. 1.0 Bark). 42 combinations of F_l and F_h were used to scan the auditory space as shown in Figure 1. The fricatives had a duration of 0.20 s and the intensity onset and offset of the natural <s> was also imitated.

A second version of the vowel context was obtained by a uniform translation of all vowel formant frequencies by + 2.5 Bark. The voice source parameters were rescaled in such a way that the mean F_0, weighted according to amplitude, was also translated by + 2.5 Bark. This transformation produces the characteristic frequencies in vowels of children four to five years of age from those of the same vowels pronounced by men <11>.

A third version of the vowel context was obtained by a uniform tonotopic compression of all formant frequencies and the weighted mean F_0. The compression is described by Equation <1>:

$$Z = Z_0 + 0.15\ (15.5 - Z_0) \qquad <1>,$$

where Z_0 is the critical band rate of a characteristic peak in the original version, and Z is the corresponding value in the compressed version. This transformation produces the characteristic frequencies of shouted vowels from those of the original <11>. Between these modes of speech, there are, however, additional differences which have not been imitated in our stimuli which provoked the impression of being produced by an aroused speaker rather than by a shouting one.

For conversion of the vowel formant frequencies f (in Hz) into critical band rate z (in Bark) Equation <2> that agrees to within ± 0.05 Bark with the empirical values <13> in the range of 0.2 to 6.7 kHz <10> was used and for reconversion Equation <3>. The formants, which were stationary, had the frequencies listed in Table 2 together with the weighted mean F_0.

$$z = (26.81\ f\ /\ (1960 + f)) - 0.53 \qquad <2>$$

$$f = 1960\ (z + 0.53)\ /\ (26.28 - z) \qquad <3>$$

Table 2: The characteristic frequencies of the three versions of the same vowels (in Hz).

	Neutral male <a>	<ɛ:>	Neutral child <a>	<ɛ:>	Aroused male <a>	<ɛ:>
F0	102	110	327	337	298	306
F1	751	442	1153	751	945	639
F2	1248	1799	1626	2617	1421	1932
F3	2501	2390	3702	3525	2558	2461
F4	3359	3413	5160	5258	3287	3332
F5	4311	4386	6977	7131	4052	4111

After D/A conversion the stimuli were recorded on tape in two different orders. First, F_l and F_h started at their highest values, 24 and 25 log. units. F_l subsequently decreased in steps of 2 u. and F_h in steps of 1 u. until the distance between the two peaks reached 7 u. In the following descending series of stimuli F_l and F_h started 1 u. below the initial values, etc. In the second order F_l and F_h started at their lowest values, 7 and 14 u., and ascended in reversal of the first order.

Each stimulus had a duration of .8 s and was presented twice in succession with an interval of 1.5 s. In the following, any sequence of this kind is considered as one "stimulus". Each stimulus was followed by a pause of 2.5 s for the subjects to respond. A pause of 5 s was inserted before each new series of stimuli. The stimuli were presented in six blocks, beginning with the neutral male version in the first (1) order, followed by child (2), aroused male (1), neutral male (2), child (1), and aroused male (2).

Procedure

The subjects were tested in a quiet, sound treated room and the stimuli were presented to them via Sennheiser HD414 headphones at a comfortable listening level. The subjects received answer sheets with a set of the five symbols "θ, s, tj, rs, sj" for each stimulus. After explaining the meaning of the symbols (<θ> or <f>, <s>, <ɕ>, <ʂ>, <ɧ>) and presenting a few stimuli for aquaintance, the subjects were asked to mark for each stimulus the symbol of the fricative they had heard. They were allowed to mark two different symbols in cases of doubt. Single-symbol responses were counted as two markings of the same symbol.

Two-dimensional histograms were obtained from the distribution of assigned labels as a function of the F_l and F_h values. The histograms were locally normalized with respect to the total number of responses to each stimulus and smoothed by a spatial cosine filter. "Phonetic boundaries", say between <s> and <ɕ>, were obtained by considering only the <s> and <ɕ> labels and computing the 50 % level curve.

RESULTS AND DISCUSSION

Effects of presentation order

"θ"-labels were infrequent and mainly attached at the highest resonance frequencies and, occasionally, at the very lowest. The boundaries between the sibilants are shown in Figure 1. The effect of contrast can clearly be seen at the <ɕ> - <ʂ> boundary which is shifted by 0.9 Bark in F_l between the two orders of presentation. Since contrast presupposes that at least one similar stimulus has been heard, there is no such effect at the beginning of each series (shown with thin lines in Figure 1). There, the responses are, instead, likely to be biased by expectation towards <s> or <ɧ> responses because the previous series of stimuli begun with these sounds. Outside this region, the <s> - <ɕ> boundary is shifted just as much as the <ɕ> - <ʂ> boundary. As for the boundary between <ɧ> and <ʂ>, the responses are likely to be biased towards <ʂ>, because this allophone would normally occur in an /aʃɛ:/ sequence as pronounced by most of our subjects. This would explain the deviant course of this boundary in the second order of presentation.

Effects of intrinsic properties

The perceptual role of the two spectral peaks in our stimuli can be understood by studying the slopes of the boundaries in Figure 1. The boundaries whose slope is not affected by order effects are well approximated by straight lines. Two of them (<ɕ> - <s> and <ʂ> - <ɧ>) have a course almost perpendicular to the F_l-axis, implying that the higher resonance F_h is practically irrelevant for these distinctions. Then, of course, the distance between the spectral peaks is also irrelevant. Thus, intrinsic properties of these stimuli were not used to normalize for speaker size.

Phonetic boundaries might possibly be given by a gross center of spectral gravity, like perceived "sharpness" <1>. Since F_h does affect the sharpness of our stimuli - as affirmed by informal listening - the results show that sharpness is not an invariant quantity in sibilants with a given phonetic quality.

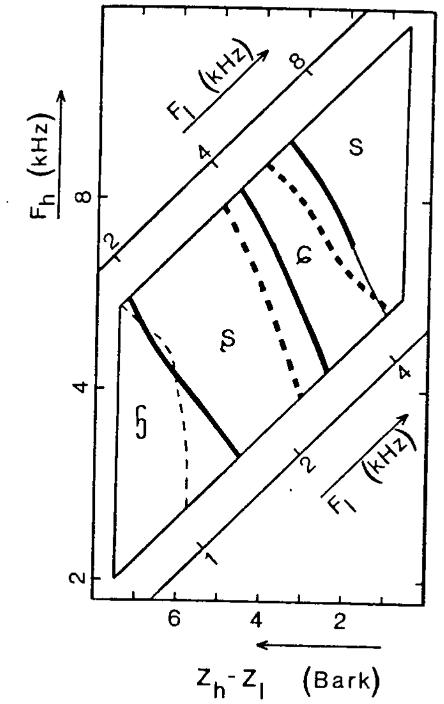


Figure 1: Phonetic boundaries between Swedish sibilants. First (continuous) and second (dashed) order of presentation. Pooled contexts.

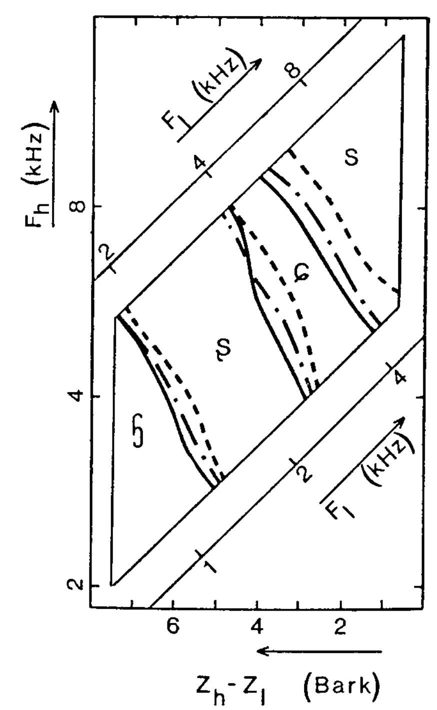


Figure 2: Phonetic boundaries between sibilants in contexts of a man's (cotinuous), a child's (dashed), and an aroused man's (dash-dotted) vowels. Pooled orders of presentation.

If the resonances are separated less than a critical distance of 3.5 Bark observed by Chistovich et al. <2> the phonetic boundaries might be expected to reflect an integrated spectral peak. The main part of our <ɕ> - <s> boundary runs through an area where $Z_h - Z_l < 3.5$ Bark (see Figure 1). The slope of this line indicates, however, that this phonetic decision is only based on the pitch of the lower spectral peak or on the spectral onset of auditory excitation. Similar results have been obtained in non-phonetic pitch matching tasks <4, 9> for frequencies below 1 kHz.

The boundaries between <ɕ> and <ʂ> are, however, not completely independent of F_h. This may be due to the fact that <ɕ> and <ʂ> are the sibilants for which our synthetic stimuli were closest to the natural versions, as judged by comparison with measured spectra of Swedish sibilants <9, 8>. The other phonetic boundaries might have followed a similar course if the stimuli had been closer imitations of natural sibilants. The phonetic boundaries can be described by Equation <4>:

$$Z_l + k_i\ Z_{hi} = I_i \qquad <4>,$$

where k_i is a factor expressing the perceptual weight of Z_{hi}, see Table 3, and I_i is a constant characteristic of boundary i. The factor k might reflect the goodness of fit between the auditory spectra of the synthetic stimuli and those of natural sibilants, but it might, alternatively, be a function of $(Z_h - Z_l)$. In that case the phonetic boundaries in Figures 1 and 2 should deviate slightly from linearity. Interestingly, k is most negative for $(Z_h - Z_l) \approx 3.5$ Bark. This reminds of the suggestion by Syrdal et al. <8> to regard this distance as specific of phoneme boundaries among sonorants. While our data do not immediately support this for sibilants - the observed boundaries are not perpendicular to the $(Z_h - Z_l)$-axis - they do show a tendency in this direction.

Table 3: Perceptual weight k of F_h in relation to that of F_l, cf. Equation <4>.

Phonetic boundary	<s>-<ɕ>	<ɕ>-<ʂ>	<ʂ>-<ɕ>	<ʂ>-<ɧ>
k	-0.05	-0.20	-0.27	-0.10

Effects of context

Since intrinsic normalization for speaker size is almost absent in our results, we would expect such a normalization, which theoretically would be appropriate, to be mediated by context. Figure 2 illustrates the effects of transforming the spectrum of the vowel context. We can see that the boundaries between sibilants are affected by the acoustic properties of the vowel context whose phonetic quality was close to invariant.

The extent of the boundary shift between the neutral male and the child version of the vowels (between +0.7 and +1.3 Bark) is, however, smaller than the translation of the vowel spectra (+2.5

Bark), especially at the <ʂ> - <ɧ> boundary. The boundaries in the aroused male version are shifted from those in the neutral version about halfway in the same direction as those in the child version. The <ʂ> - <ɧ> boundary (at 11.6 Bark = 1.6 kHz) is shifted by roughly +0.3 Bark, i. e., less than the vowel formants in the same frequency region (+0.6 Bark). Since, further, the upper vowel formants (above 15.5 Bark = 2.9 kHz) in the aroused male version are not shifted upwards but slightly downwards, the shift of the <s> - <ɕ> boundary (at Z_1 = 19 Bark) can not have been guided by the vowel formants in the same frequency region. Apparently, the sibilant boundaries are shifted about half as much as some weighted mean of the vowel formants, F2 given the highest weight. This would hold approximately for both of our context transformations, but the correlation of the extent of boundary shift with F_1 remains an open question.

ACKNOWLEDGEMENT

This research has been supported by a grant from HSFR, the Council for Research in the Humanities and Social Sciences.

REFERENCES

<1> G. v. Bismarck, Extraktion und Messung von Merkmalen der Klangfarbenwahrnehmung stationärer Schalle, München 1972.

<2> L. Chistovich and V. Lublinskaya, "The "center of gravity" effect in vowel spectra and the critical distance between formants", Hearing Res. 1, 1981, 185-195.

<3> G. Fant, "Glottal source and excitation analysis", STL-QPSR 1/1979, 85-107.

<4> R. Glave Untersuchungen zur Tonhöhenwahrnehmung stochastischer Schallsignale, Helmut Buske Verlag, Hamburg, 1973.

<5> J. M. Heinz and K. Stevens, "On the properties of voiceless fricative consonants", J. Acoust. Soc. Am. 33, 1961, 589-596.

<6> P. Lindblad, Svenskans sje- och tje-ljud i ett allmänfonetiskt perspektiv, CWK Gleerup, Lund 1980.

<7> J. Martony, "On the synthesis and perception of voiceless fricatives", STL-QPSR 1/1962, 17-22.

<8> A. K. Syrdal and H. S. Gopal, "A perceptual model of vowel recognition", J. Acuost. Soc. Am. 79, 1986, 1086-1110.

<9> H. Traunmüller, "Perception of timbre: ", in R. Carlson and B. Granström (eds.), The Representation of Speech in the Peripheral Auditory System, Elsevier Biomed., 1982, pp. 103-108.

<10> H. Traunmüller, "Analytical expressions for the tonotopical sensory scale", part of Ph. D. thesis, Stockholms Universitet, 1983.

<11> H. Traunmüller, "Some aspects of the sound of speech sounds", contr. to NATO-ARW on psychophysics of speech perception, Utrecht 1986.

<12> H. Traunmüller and F. Lacerda, "Perceptual relativity in identification of two-formant vowels", Speech Communication 5, 1987, ...

<13> E. Zwicker, "Zur Unterteilung des hörbaren Frequenzbereiches in Frequenzgruppen", Acustica 10, 1960, p. 185.

ASPIRATED VS. NONASPIRATED STOPS AND AFFRICATES IN STANDARD CHINESE

Zong-ji Wu and Yi Xu

Institute of Linguistics
Chinese Academy of Social Sciences
Beijing, China

ABSTRACT

The distinction between aspirated and non-aspirated consonants in Standard Chinese (SC) is usually described in traditional phonetics as a difference in force or glottis opening. Our experiments, including acoustic analysis, manometers measurements and perception tests with synthesized consonants, revealed that information about aspiration is carried by prolonged turbulence with different features. The perceptive cues for aspiration in affricates also depend upon the tongue positions of the following vowels: when before a low vowel, the aspiration is realized as a fricative /h/ immediatly following the releasing noise, while before a high vowel, it is realized as the prolongation of the releasing noise.

INTRODUCTION

In Standard Chinese, there are two groups of consonants which can be produced both with and without aspiration. These are the voiceless stops and affricates, each of which is distinguished from its counterpart in the feature aspirated/non-aspirated in the Chinese phonology.

In many European languages, e.g. English, the aspiration of word initial stops is only a conditional feature, but in many tone languages, especially in Chinese, it is a phonemic feature. In traditional Chinese phonetic works, the nature of aspiration is mostly discribed in terms of the force of articulation: the aspirated consonants having greater force of articulation than the non-aspirated ones. One of the popular phonetic outline books stated, "In aspirated articulation, the air stream expelled from the mouth cavity is stronger than in non-aspirated articulation.[1] It is also mentioned here and there that "an air-flow after release is called aspiration", that "the air-flow in non-aspirated sound is weaker and shorter, and vise versa, and that "for an aspirated sound, the glottis is opened during release, the air pressure is large and the air-flow breathing out is obvious", etc.

In recent decades, thanks to the widely application of phonetic experimentation, phoneticians can study the problems of aspiration more deeply, and the veil of non-aspirated/aspirated distinction are now being raised gradually. Many technics have been used for investigating this feature in the levels of articulation, acoustics and perception. The VOT features of aspirated/ non-aspirated oppositon was examined in spectrograms and was proved by perception tests as an important cue[2], there were also investigators dealing with the glottic movements, air-flow rates, and nerves activities. Those studies have brought the discussion of aspiration to a high level.[3]

This paper intends to make further studies on the non-aspirated/aspirated consonants in Standard Chinese in order to raise and answer the following questions:

1) Which is the main perceptive cue for aspiration, the air-stream force, the duration or VOT, or the glottis opening?

2) What are the articulatory processes of these consonants?

3) Are there any different aspiration features between stops and affricates?

EXPERIMENT

Early in 60's we tested all the SC consonants, spoken by two speakers, a male and a female, of the Beijing dialect, using a level recorder (type: BK 2304) to measure the amplitudes and the length of the consonantal segments. The amplitude represents the overall acoustic pressure and the length was measured from the release point to the starting point of vowels.

For measuring the concentration area of noise and the VOT as well as the transition cue in stops and affricates, a Kay sonagraph of model 7029 was used. The amplitude were measured immediately after the release. The materials were spoken by a male Beijing native.[4]

Two sets of manometers were used to

measure the supra-glottis air-pressures and air-flow rates, the equipments were constructed by Professor Peter Ladefoged. [5] Thanks to the Department of linguistics at UCLA, the experiments were done in their lab by Dr.H.M. Ren. Two informants, a male and a female, were asked to pronounce all the stops and the affricates in Standard Chinese, each followed by three vowels, high or low.

For the perception test, a number of non-aspirated and aspirated stops and affricates were synthesized by a synthetic system designed by the phonetics laboratory of the Institute of Linguistics, Chinese Academy of Social Sciences.[6] Selected samples of spectrograms were made of the synthesized syllables.

For comparison, we made several spectrograms from Miao language in Guizhou and Bai-ma language in Tibet, which have aspirated fricatives in contrast with non-aspirated fricatives, in order to examine the nature of the aspiration noise. The materials were kindly supplied by the Institute of Nationalities, Chinese Academy of Social Sciences.

DISCUSSION

FORCE OR DURATION? To determine whether the force or the length of the noise plays the major role as the perceptive cues for aspiration, a number of experimental techniques were used. As the space of the present paper is limited, only a few selected examples are given here. Fig.1 and Fig.2 are histograms of the amplitudes measured from the acoustic records spoken by two subjects A and B (a male and a female). The amplitude of aspirated stops and affricates are shown somewhat stronger than that of unaspirated ones, especially in /pa/ and /pu/, where the explosion of unaspirated stops are too

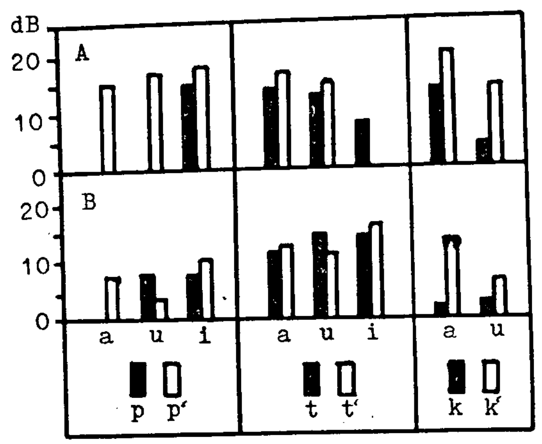


Fig.1 Histograms of the amplitude of non-aspirated and aspirated stops

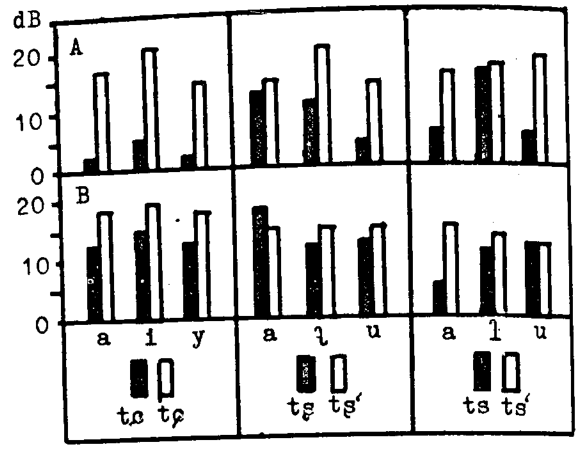


Fig.2 Histograms of the amplitude of non-aspirated and aspirated affricates

weak to be detected. But in the /p'u/, /t'u/, /ts'a/ and /ts'u/ spoken by B, the results turned out to be just in the contrary. As a whole, the difference in force between the aspirated and non-aspirated consonants are not so evident as commonly believed. In Fig.3 and Fig.4, great differences can be seen in the measurements between the length of the aspirated and non-aspirated consonants; that for the stops, the proportion of amplitude in the aspirated and non-aspirated stops is 11/7 and that of duration is 71/11, while for the affricates the proportion of amplitude in the two categories is 14/10 and the proportion of duration is 120/46. On the average, their differences are around 1.5 to 1 in amplitude and 3 or more to 1 in duration.

Table I gives the data of supraglottal air-pressure of both stops and affricates. There are no direct proportional relationship between aspirated and

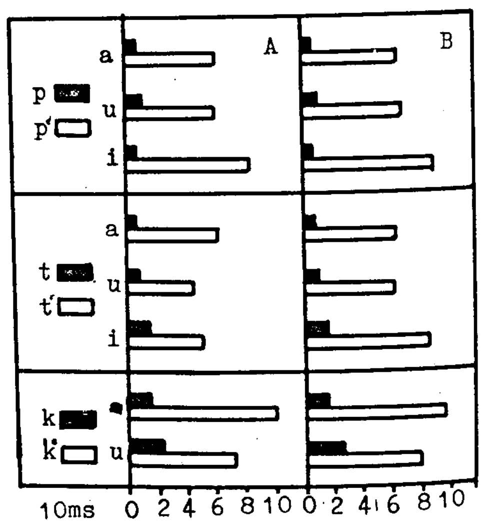


Fig.3 Histograms of the duration of non-aspirated and aspirated stops

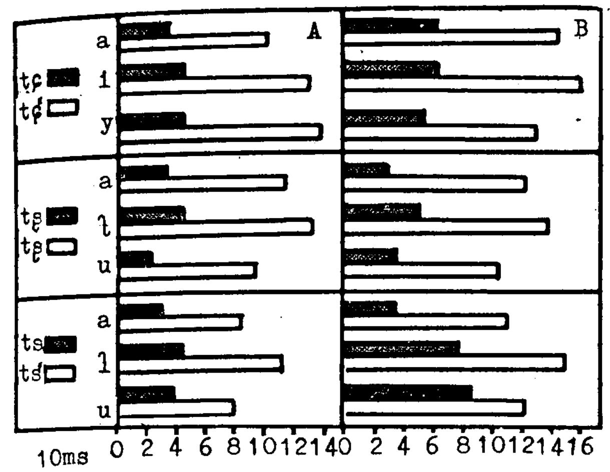


Fig.4 Histograms of the duration of non-aspirated and aspirated affricates

Table I Supraglottal air-pressure of stops and affricates in SC

		p	t	k	ts	tʂ	tɕ
Non-asp	A	10	90	-	11	89	14
	B	10	27	-	11	31	58
		pʻ	tʻ	kʻ	tsʻ	tʂʻ	tɕʻ
Asp	A	11	84	3	11	39	17
	B	10	24	-	19	10	25

Table II Supraglottal air-flow rate of stops and affricates in SC

		p	t	k	ts	tʂ	tɕ
Non-asp	A	-	54	-	50	-	-
	B	-	-	-	-	-	-
		pʻ	tʻ	kʻ	tsʻ	tʂʻ	tɕʻ
Asp	A	90	100	110	100	120	80
	B	100	50	90	100	120	80

non-aspirated consonants. Moreover, the proportion of pressure in /t'/ and /t/ is 84/90 and in /ts'/ and /ts/ is 39/89 by speaker A, and in /tɕ'/and /tɕ/ is 25/58 by B. The air-pressure of aspirated consonants is occasionally weeker rather than stronger than that of non-aspirated ones. Table II gives the data of supraglottal air-flow rates, which are directly proportional to the duration measured in the acoustic dimension. That is to say, the more the air-flow, the more the noise is prolonged. The difference in supraglottal air-flow between aspirated and non-aspirated consonants are much greater. It is easy to explain. In unaspirated stops, the explosive noise is short and is immediately followed by the vowel. There is occasionally a silent gap between CV instead of a noise. In unaspirated affricates, although the friction seems to be the results of air-flow, the air does not expel continuously out from the pulmonic cavity while the glottis keeps open. Mowover, the width of the glottis opening is much more smaller than that of aspirated affricates [7]. So their air-flow rate could hardly be measured.

SOUND SOURSE The aspiration of consonant is usually defined as a noise. On the acoustic point of view, questions might be raised as: what are characteristics of these noise? Are they all white-noise, or noise with different parameters?

In the spectrograms of these consonants, in an aspirated stop as /p'/, the aspirated section is a sequence of non-periodical noise, bearing the acoustic features similar to that of fricative /h/. i.e., the concentration areas are scattered and connected the formants of the following vowel with a /plain/ transitional feature. While in an aspirated affricates as /ts'a/, the sequence of noise is separated into two: the first part of the noise bears the same acoustic character as the fricative /s/, and the second part is the same as /h/. Thus a clear boundary between them is shown. But when an aspirated affricated followed by another vowel as in /ts'ɿ/, the noise is quite different from that in /ts'a/. It gives no /h/ noise but prolongs the /s/ friction. The same phenomina can be also found in /tʂ'ʅ/ and /tɕ'i / in Standard Chinese. These can be explained by the physiological interpretation. When the vowel after an aspirated affricate, if the tongue height is lower than that of the consonant as in /ts'a/, as soon as the constriction of /ts/ is released and the tongue moves to /a/, no turbulence will be produced with the tongue tip, then the aspiration has to be formed by another way, so a /h/ like turbulence is produced at the back area of the tongue. This can be seen in the X-ray films. But in /ts'ɿ/, when the vowel starts, the tongue position does not move far apart from the upper palate for the gesture of the vowel is homorganic with that of /ts' /, but the stricture is slightly enlarged and the turbulence is displayed by the voice.

PERCEPTION TEST In order to prove the results mentioned above, a number of perception tests are arranged through a synthesizer. Some of the consonants given in Fig.1-4 are sythesized by rule in which the aspirated sections are changed by couple the amplitude or couple the length of the friction. The parameters of the frictive part of the affricates are given based on the quality of /h/ or of the same as the friction of affricates. Fig.5 is a sample of an affricate /tʂ/ followed by an open vowel /a/. From left to right, /tʂ/ is a non-affricate, /tʂ/

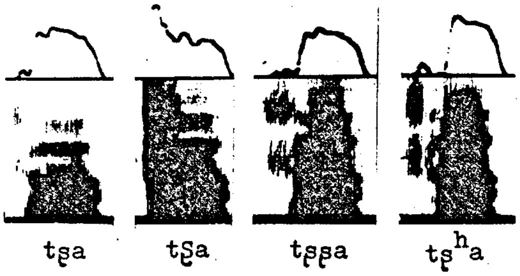


Fig.5 Spectrograms of synthesized affricates /tʂ/ in different nammers followed by vowel /a/ (see text).

double the amplitude of /ʂ/, /tʂʂ/ double the length of /ʂ/ and /tʂh/, is /ts/ followd by /h/. The perception result is promissing in the last sample, and in which a boundary between the two frictions is prominantly seen. Fig.6 is a sample of affricate /tʂ/ followed by the vowel /ʅ/, and a /tɕ/ followed by the vowel /i/, both with their frictions doubled the length or followed by /h/; better results are obtained by doubled the lengths instead of plus a /h/.

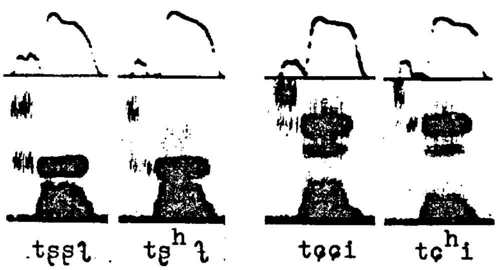


Fig.6 Spectrograms of synthesized affricates /tʂ/ and /tɕ/ in different manners followed by vowels /ʅ/ and /i/ respectively (see text).

It is interesting to have this results revealed in certain minority languages in China. For example, there are non-aspirated/aspirated pairs in Miao language of Guizhou, both the affricates and frictives can be aspirated. Fig.7 shows two pairs of "sa"/"s'a" and "ɕi"/"ɕ'i", in which we can see that "s'a" is "s" plus "h" with a boundary in the friction, while "ɕ'i" is a prolonged "ɕi" without any boundary in the friction.

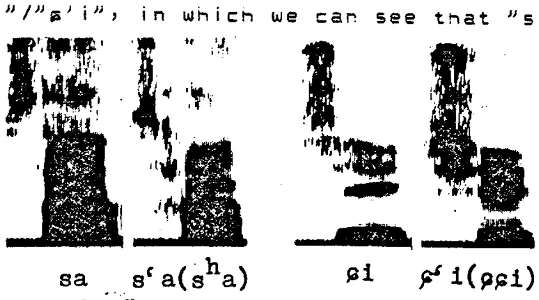


Fig.7 Spectrograms of non-aspirated and aspirated fricatives /s/ and /ɕ/ in Miao language.

CONCLUSION

The stops and the affricates in Standard Chinese exist two manners of articulation, non-aspirated and aspirated. The perception cues are mostly based upon the noise duration rather than the force. Moreover, the acoustic features of the aspirated noise are different in two types according to the following vowels. The aspiration is formed by adding a /h/ sound after release if it is followed by an open vowel; while formed by prolonging the length of noise if followed by a high vowel homorganic with the consonant.

REFERENCE

[1] C.P. Luo, J. Wang, Outline of General Phonetics, Commercial Press, 1981.
[2] A.S. Abramson, L.Lisker, Voice-timing perception in Spanish word initial stops, J. Phonetics 1, 1973
[3] H. Hirose, Laryngeal adjustment in consonant production, Phonetica 34, 1977
[4] Z.J. Wu ed., The Spectrographic Album of Mono-Syllables of Standard Chinese, C.S.S.P., 1986.
[5] P. Ladefoged, A. Traill, Instrumental phonetic fieldwork, Report in Insti. Ling. Chi. Sci., 1983.
[6] S. Yang and Y. Xu, A software system for synthesizing Chinese speech, Proc. Inter. Cong. on Chi. Infor. Process, Beijing, 1987.
[7] R. Iwata, H. Hirose, Fiberoptic acoustic studies of Mandarin stops and affricates, Ann. Bull. RILP 10, 1976.

PERZEPTIVE BEWERTUNG DER TSCHECHISCHEN EXPLOSIVLAUTE

Sáva Hlaváč

Phonetisches Labor
Institut für tschechischen Sprache ČSAV
Praha, ČSSR

RESUME

Die beschriebenen Experimente sollten Beziehungen zwischen signifikanten Merkmalen der tschechischen Explosivlaute feststellen. Der Wichtigkeitsgrad, gemeinsame Vertauschmöglichkeit und Reduktion der Merkmale wurden definiert. Die Merkmale waren: das explosive Geräusch, das postexplosive (PE) homorgane Geräusch und Transiente an dem die Konsonante begleitenden periodischen Signal.

EINLEITUNG

Gegenwärtige Methoden der Sprachsynthese benutzen meistens digitalisierte Segmente mit der Länge 20 + 30 ms. Im Verarbeitungsprozeß muss man wichtige perzeptionsrelevante Merkmale bestimmen, die keine Verzerrung erleiden dürfen. Für Explosive ist das besonders wichtig ihres transienten Charakters wegen. Das Sprechsignal wird durch verzerrende Einflüsse beeinträchtigt; zugleich wirkt aber auch deren Kompensationsfaktor. Die Gesamtverständlichkeit kann auch dann erhalten bleiben, wenn einzelne Faktoren der Laute in veränderter Weise realisiert werden. In der Sprachsynthese kommt es darauf an, welcher Teil des akustischen Signals für die Perzeption des untersuchten Explosivs der wichtigste ist: das explosive Geräusch, das PE homorgane Geräusch oder Transiente an dem die Konsonanten begleitenden periodischen Signal. Das Experiment wurde mit tschechischen stimmlosen Explosivlauten [p], [t], [t1], [k] durchgeführt. Bei der Bewertung anderer Explosivlaute muß man die Verschiedenheiten in der Lautbildung und deren perzeptionellen Bewertung respektieren.

BENUTZTE MESSMETHODE

In der Sprechforschung benutzt man häufig die segmentale Synthese. Für die Analyse der Sprechelemente ist es jedoch zweckmässig mit natürlichem Signal zu arbeiten. In unserem Labor verwendeten wir die Methode der mechanischen Montage des bei höher Geschwindigkeit 76,2 cm/s arbeitenden vollspurigen Tonbandes. Mittels dieser Methode ist es möglich:

1. beliebige Kombination der Signalabschnitte zu konstruieren und einzelne Teile dabei auslassen
2. ausgewählte Abschnitte genau an Stellen zu bringen die vorher im Spektrogramm definiert wurden
3. voraus- definierte Signalübertragung zu bestimmen und das Signal in der Intensität zu modifizieren
4. die gesamte Aufnahmelänge konstant zu halten bzw. sie nach Bedarf zu modifizieren
5. einzelne Teile des Signals durch andere zu ersetzen
6. das Originalspektrum der Restteile und die Zeitverhältnisse nur minimal beeinträchtigen.

Die Zeitdauer des Überganges der verbundenen Teile war meistens 10 ms (Abb. 1).

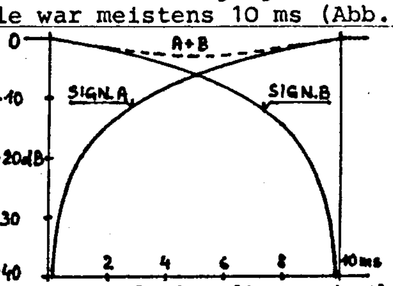


Abb. 1

Diese Länge ermöglichte die maximale Information über das ursprüngliche Signal zu behalten und gleichzeitig das Entstehen von Übergangserscheinungen zu begrenzen. Der Übergang wurde dort gelegt, wo konnte man keine signifikante Signaländerung erwarten.

Das Material wurde wie folgt verarbeitet:

a) Das Spektrogramm der Aufnahme wurde verfertigt.
b) Nach Auswerten des Spektrogramms wurde mittels Montage ein modifiziertes Signalmuster erstellt.
c) Das Spektrogramm des Musters wurde verfertigt.
d) Das Muster wurde perzeptiv gewertet.

Die Bewertung der Ergebnisse ist dadurch beeinflußt, daß das Maß an Wichtigkeit des perzipierten Signals in keinem konstanten, linearen Verhältnis zur Intensität steht. Die genaue Auswertung der Zeit- und Spektralverhältnisse muß also sehr sorgfältig ausgeführt werden.

GRAPHISCHE DARSTELLUNG UND PHYSIOLOGISCHE PROZESSEN

Unter Anwendung der Booleschen Algebra können die, für die Identifizierung des tschechischen Explosivs notwendigen Bedingungen wie folgt definiert werden

$$E_n = \bar{F}_o \; F_{char} \; T$$

E_n - Erkennbarkeit des nicht stimmhaften Explosivs
F_o - Grundton der Stimme
F_{char} - charakteristisch verstärkte Frequenzbereiche
T - Zeitfaktor (Dauer, Änderungsgeschwindigkeit...)

Eine Analyse des Artikulationsprozesses zeigt, daß nach einer relativ langen Unterbrechungsphase des Vokaltraktes eine jähe Öffnung folgt. Ihrer phasikalischen Realisierung entspricht ein Ablauf der nur einen Teil der mit $k_1(1 - \exp(-k_2t))$ ausgedrückten allgemeinen Exponentialkurve bildet. Das Öffnen des Stimmweges verläuft nur in erster Approximation sprunghaft, der Erregungsimpuls entsteht also mit beschränkter Geschwindigkeit. Das breiteste Spektrum erregt der Anlaufteil des Impulses, sofern dieser in genügender Intensität entsteht: dies wird zum größten Teil durch die Artikulationsweise beeinflußt. In der weiteren Phase des Öffnens entsteht ein homorganes postexplosives Reibegeräusch, modifiziert durch das Erweitern des Spaltes. Beide Geräuscharten werden durch die filtrierende Wirkung der Resonanzhöhlen beeinflußt. Auch die Parameter der Höhlen sind veränderlich, meistens zugleich mit Veränderungen des Geräusches. Folgt auf ein Explosiv ein weiteres Explosiv, endet der ganze Prozeß bei der dem zweiten Explosiv entsprechenden Okklusion. Das Geräuschsignal des ersten Explosivs ist so zeitlich begrenzt. Folgt ein Vokal, schließt sich verhältnismäßig bald nach der Explosion ein periodisches Signal an. Durch Resonanz wird dieses Signal in Frequenzbereichen verstärkt, die durch das ansonst abschwächende homorgane Geräusch des ausklingenden Explosivs betont werden. Das PE Geräusch wird kontinuierlich abgeschwächt und durch das periodische Signal maskiert; in seiner Anfangsphase kann also dieses zur Erkennbarkeit des Explosivs beitragen. Der letzte Teil des fließenden Überganges gehört zum typischen Bereich des folgenden Vokals. Beim Anhören des isolierten Teils aus dem mittleren Bereich des Überganges kann weder die Zugehörigkeit zum Vokal, noch diejenige zum Explosiv mit genügender Genauigkeit bestimmt werden.
Bei guter Aussprache sind im Spektrum zwei Teile zu erkennen: die Explosion und das postexplosive Geräusch, bei verschlechterter Qualität der Aussprache verschlechtern sich beide gleichzeitig. Bei geläufigem Sprechen ist die Explosion häufig geschwächt, oder (in spektraler Darstellung) fehlt sie ganz. Im Falle der segmentalen Darstellung des Sprechsignals (z.B. bei der LPC Synthese) kann die Explosion entweder ganz fehlen oder wird sie in der Länge des ganzen Segments dargestellt. Solche Weglassung oder Verlängerung wirkt sich meistens als störend aus.

EXPERIMENTELLE ANORDNUNG DER VERSUCHE

Die durchgeführten Experimente sollten folgende Fragen beantworten:
1. Welche ist die Bedeutung der eigentlichen Explosion und des PE Geräusches?
2. Welche sind die Folgen einer Verkürzung der eigentlichen Explosion bei fehlendem PE Geräusch?
3. Welche Folgen hat das Aussetzen der Explosion bei behaltenem PE Geräusch?
4. Welche sind die Folgen der fehlenden Explosion und des PE Geräusches bei behaltenem periodischen Anlaufteil des folgenden Vokals?
5. Welchen Einfluß auf die Perzeption des Explosivs hat das vollkommene Beseitigen des periodischen Anlaufteils des Vokals bei behaltener Explosion?
6. Gelten analogische Schlüsse, wenn das Explosiv vor einem anderen Explosiv oder einem Vokal steht?

Die Versuche wurden mit tschechischen Wörtern entsprechend den Formeln C_1C_2V und $C_1C_2C_3V$ durchgeführt. Untersucht wurde der Einfluß von Veränderungen am
a) explosiv in der Initialstellung (C_1)
b) explosiv in der Medialstellung (C_2 vor C_3)
c) explosiv in der Medialstellung vor dem Vokal (C_2V)
d) Übergangsgebiet des Vokalanlaufs nach einem Explosiv (C_2V)

Die Worte wurden aus Tonbandaufnahmen guter Aussprache in einer akustisch gedämpften Kammer des Phonetischen Labors des Instituts für tschechischen Sprache der ČSAV in Praha ausgewählt. Die Aufnahme wurde mit 2 männlichen Stimmen durchgeführt. Proben normaler Aussprache wurden nach den Regeln der zufälligen Anordnung mit mittels Montage modifizierten Proben vermischt, und ihre Perzeption wurde mittels einfacher Anhörteste überprüft. Bei der Probenverfertigung benutzte man ein Eichsignal 2 kHz, 100 ms, im Abstand 100 ms vor dem Probeanfang. In jedem Test wurden Proben derselben Stimme vergleichen. Die Anhörgruppe betrug etwa 10 Personen (7 + 11); es handelte sich dabei um geborene Tschechen mit phonetischer Ausbildung sowie auch ohne dieser, jedenfalls mit normalem Gehör. Gewertet wurde nach einer dreistufigen Skala:
1. Normaler Gehöreindruck, ohne wahrnehmbare Veränderungen
2. Gehöreindruck mit wahrnehmbaren Qualitätsänderungen
3. Erkennbarkeit des Lautes gleich Null (der Laut nicht perzipiert, durch anderen Laut substituiert, nicht erkennbar).

Das Auswerten der Ergebnisse geschah nach Schema: Urteile (1 + 2) gegenüber 3.

Die Explosive nach a) und b) wurden in zwei Tests (A, B) verarbeitet. Die mittels Montage modifizierten Proben enthielten in diesen Fällen sowohl das Auslassen der Explosion (-E) als auch das Auslassen des homorganen PE Geräusches (-PE).

Die Explosive nach c) und der Übergang nach d) wurden in vier Tests verarbeitet (C, D, E, F). Im Test C wurden die Proben des normalen (V_n) oder veränderten (V_{med}) Vokals kombiniert mit Modifikationen des Explosivs mit ausgelassener Explosion (-E) oder ausgelassenem PE Geräusch (-PE). Die veränderte Probe (V_{med}) entstand dadurch, daß der normale Anlaufteil der periodischen Schwingung des Vokals durch den medialen - stationären - Teil des Vokals ersetzt wurde. Es entstanden 4 Modifikationen: $-EV_n$, $-PEV_n$, $-EV_{med}$, $-PEV_{med}$.

Der Test D enthielt Modifikationen der Länge des explosiven Geräusches vor dem Vokal mit ersetztem Anlaufteil. Der Rest des PE Geräusches war 5 ms und die Explosion wurde auf die Dauer von 5, 10 und 15 ms begrenzt. Die Proben $E_{15+5}V_{med}$, $E_{10+5}V_{med}$, $E_{5+5}V_{med}$, $E_{0+5}V_{med}$.

Der Test E bestand aus Proben mit ausgelassener Explosion und ohne PE Geräusch, in denen entweder der normale Vokalanlauf folgte, oder war der Anlauf durch den Medialteil ersetzt. Ein langsamer Intensitätsanlauf dieses substituierten Teils erfolgte mit Längen von 10, 20, 40, 70 ms: so wurde das weiche Ansetzen der Intensität substituiert, bei dem die Komponente der Frequenzveränderung fehlte (die aber im normalen Sprechsignal immer vorhanden ist). So entstanden die Proben $-EV_n$, $-EV_{med10}$, $-EV_{med20}$, $-EV_{med40}$, $-EV_{med70}$.

Im Test F wurden Proben mit Explosion vor dem Vokal, dessen Anlaufteil durch den Medialteil ersetzt war, mit Proben ohne Explosion und PE Geräusch vor dem Normalvokal, oder vor dem ersetzten Anlaufteil kombiniert. Proben $+EV_{med}$, $-EV_n$, $-EV_{med}$.

Der angeführte Arbeitsgang ist anhand ausgewählter Beispiele dokumentiert. Die Abbildungen zeigen Spektrogramme des tschechischen Wortes mit Änderungen in Position nach c) und d) und Verarbeitung nach E, F. Es handelt sich um das Wort [tka:t] (≡ weben). Abb. 2 zeigt das Originalspektrum mit gut entwickeltem Übergang der Laute [ka:].

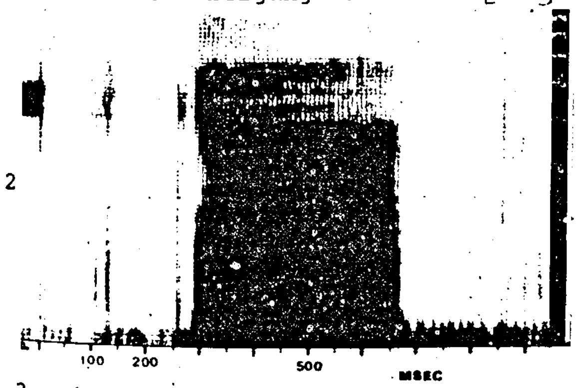


Abb. 2

Abb. 3 zeigt dasselbe Spektrum mit ausgelassener Explosion und ohne PE Geräusch.

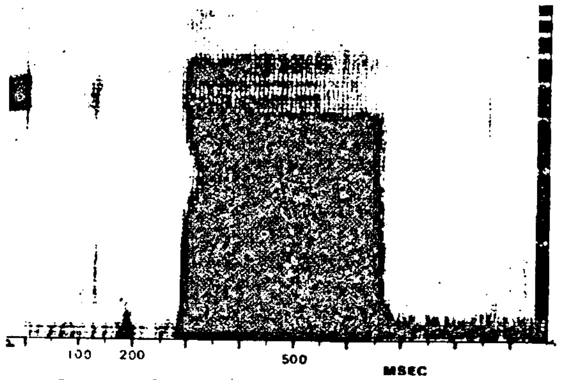


Abb. 3

Abb. 4 zeigt ein Spektrum mit behaltener Explosion und mit PE Geräusch. Der Anlaufteil ist durch Signal aus dem charakteristischen Medialbereich des folgenden Vokals [a:] ersetzt.

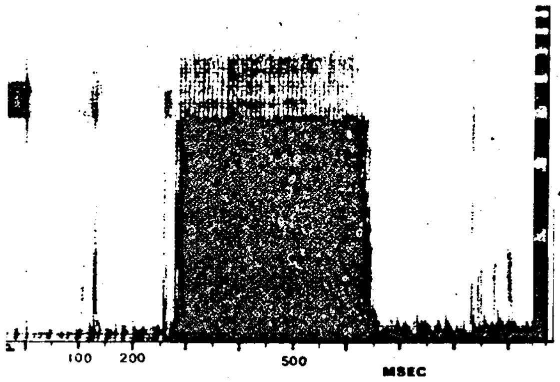


Abb. 4

Abb. 5 zeigt das Spektrum mit ausgelassener Explosion und PE Geräusch und mit ersetztem Anlaufteil des Vokals (ähnlich wie in Abb. 4). Die Länge der Proben entspricht der Originalaufnahme. Bei der Auslassung des Originalsignals wurde ein Blank derselben Länge aus demselben Tonband benutzt. Partien ohne Sprechsignal sind mit gleichem Geräuschhintergrund der Tonaufnahme wie beim umgebenden Signal dargestellt. Die Montage nach Abb. 4 und 5 kann man künstlich (nur im Labor) erzielen. In der Sprache ist es infolge der endlichen Trägheit der sich bewegenden Stimmorgane nicht erreichbar. Zur Kontrolle diente die Anordnung nach Abb. 5, in der der Einfluß aller Signale des Explosivs vollkommen unterdrückt wurde. Ähnliche Verarbeitung wurde auch für den Fall mit dem Explosiv in der Medialstellung benutzt (C_2 in der Gruppe $C_1C_2C_3V$).

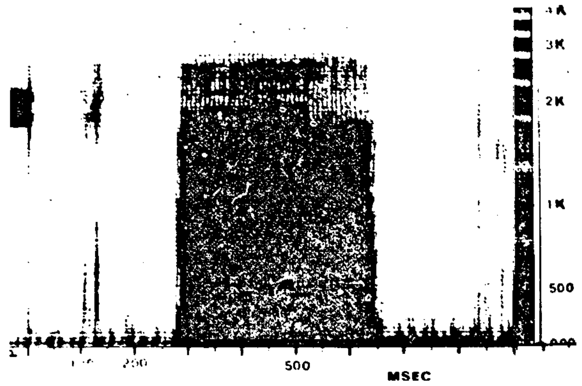


Abb. 5

Bei der Verarbeitung der Ergebnisse benutzte man statistisches Testverfahren. Jede Probe wurde im Test 2 + 3 mal präsentiert. In jedem Test wurde das arithmetische Mittel der richtigen Antworten und die Standardabweichung ermittelt. Richtige Antworten bewegten sich im Bereich ungefähr 83 + 100%.

AUSWERTEN DER ANHÖRTESTE

Das Auswerten der Teste ergab:

1. Die Perzeption des Explosivs in Medialstellung blieb im Prinzip ungestört:
   a) bei Auslassen der Explosion, wenn ein PE Geräusch folgte,
   b) bei Auslassen des PE Geräusches, wenn die Originalexplosion vorhanden war.
2. Wurde das PE Geräusch ausgelassen, führte ein weiteres Kürzen der Explosion zu einer schlechteren Erkennbarkeit des untersuchten Explosivs. Im Grenzfall verschwand es vollkommen. Es handelte sich dabei um eine fehlende Perzeptionserscheinung, nicht um das Verwechseln mit einem anderen Laut.
3. Wurde im Signal nur das PE Geräusch behalten, blieb auch in diesem Falle die Perzeption weitgehend ungestört.
4. Wurden sowohl Explosion als auch PE Geräusch ausgelassen und der Anlaufteil des folgenden Vokals behalten (wie in Abb. 3), blieb die Perzeption ebenfalls ungestört.
5. Das Vertauschen des periodischen Anlaufteils des Vokals bei behaltener originaler Explosion (Abb. 4) hat auf die Explosivperzeption ebenfalls keinen bedeutenden Einfluß gehabt.
6. Analogische Schlußfolgerungen zeigen, daß jedes der drei untersuchten Elemente für die Perzeptionserscheinung allein genügt. Die Kombination von zwei oder allen Elementen hat redundanten Charakter und verbessert die Verläßlichkeit der Perzeption.

Im Kontrollversuch wurde die Explosion als auch das PE Geräusch ausgelassen und der periodische Anlaufteil des Vokals wurde durch den Medialteil ersetzt (Abb. 5). Das relevante Explosiv konnte keinesfalls identifiziert werden. Die ursprüngliche Gesamtqualität blieb dabei erhalten. Dies führte zu einer merklichen Verlängerung der Vokalperzeption. Bei Verlängerung des Intensitätsanlaufs der substituierten periodischen Komponente in Abb. 4 und 5 (schrittweise bis 80 ms) konnte die Voraussetzung nicht bestätig werden, daß ein hartes (kürzeres) Ansetzen des periodischen Teils die Identifizierung des Explosives verbessert, sofern es nicht zugleich von einer entsprechenden Frequenzänderung begleitet wird.

Aus den Messungen folgt ebenfalls der Beweis der hohen Perzeptionempfindlichkeit des Gehörs für signifikante Übergangsprozesse. Z.B. eine Verkürzung des relevanten Geräuschsignals des [k] bis auf ca 15 + 10 ms (Extremfall) hatte keine Einfluß auf die Erkennbarkeit des Explosivs, unter der Voraussetzung einer Explosion mit gut entwickeltem Spektrum.

Im natürlichen Sprechsignal ergänzen sich also gegenseitig alle Komponenten: Explosion + PE Geräusch event. die periodische Komponente im Anlaufteil des folgenden Vokals. Bei idealer Artikulation ist das nach der Explosion folgende Signal redundant, umgekehrt bei schlechter Artikulation mit fehlender (ungenügend entwickelter) Explosion kann die zweite oder die dritte Komponente die Explosion vertreten. Besonders markant ist diese Tatsache bei der Kombination Explosiv + Vokal. Diese Wertung ist natürlich diskutabel: es kann ebenso behauptet werden, daß im geläufigen Sprechsignal der Einfluß der Übergangsgebiete primär zur Geltung kommt, während die eigentliche Explosion eher zur redundanten Information gehört.

SCHLUSSFOLGERUNG

Die beschriebene Methode demonstriert die Annahme, daß tschechische Explosive mittels mehrerer sich gegenseitig vertretender Merkmale bestimmt werden können. Die Spektrogramme der modifizierten Signale wurden zur Auswertung des Charakters von mittels der Perzeptionsteste identifizierten Proben benutzt.
Die Perzeptionsauswertung der tschechischen Explosivlaute Beruht also auf der Bewertung der einander sich vertretenden Merkmale (Redundanzprinzip). Für richtige Bewertung des Explosives genügt nur ein einziges Merkmal von den hier beschriebenen (Einzelfälle natürlicher und künstlicher Sprachdegradation). Schlußfolgerungen kann man bei der Sprachsyntheseprogrammierung ausnutzen.

LITERATUR

1. F.S.Cooper, P.C.Delattre, A.M.Liberman, J.M.Borst, L.J.Gerstman: Some experiments on the perception of synthetic speech sounds. JASA 24, 1952, 597-606.
2. E.Fischer Jörgensen: Tape cutting experiments with Danish stop consonants in initial position. ARIPUC 6, 1972, 104-- 168.
3. A. van Katwijk, J. t'Hart: Intelligibility of syllable - tied interrupted speech. I.P.O. Report April 1967, 99-- 102.
4. B.Lindblom: Accuracy and limitations of Sonagraph measurements. Proc. of the IV Int. Congr. of Phon. Sci., Helsinki 1961.
5. G.E.Peterson, W.S.Wang: Segmentation techniques in speech synthesis. JASA 30, 1958, 739 - 742.
6. G.Ungeheuer: Systematische Signaldestruktion als Methode der psychoakustischen Phonetik. Phonetica 18, 129.

AUTOMATIC ASSESSMENT OF MACHINE TRANSCRIPTIONS

Peter Roach   Andrea Dew   Paul Rowlands

Department of Linguistics & Phonetics
University of Leeds, U.K.

ABSTRACT

Research in the automatic transcription of speech sounds by computer requires a detailed and accurate comparison between the expert phonetician's transcription and the machine's attempt. A computational technique for assessing the accuracy of a machine transcription is described: differences between segments are expressed in terms of a small number of primitive phonetic features.

INTRODUCTION

A number of modern approaches to the automatic recognition of continuous speech make use of the technique of dividing the stream of speech into a string of segments and labelling these with a chosen set of phonetic category labels ([1], [2], [3], [4], [5]). These categories, which are not necessarily restricted to phoneme-sized units, may be more or less precisely specified. Dalby et al [6] refer to three different types of analysis: Broad Class (identifying segments as, for example, Nasal, Fricative, Vowel), Mid Class (including details such as whether a segment is voiced or not, whether a vowel is front or back, or whether a fricative is strong or weak), and Fine Class, which is roughly equivalent in precision to a phonemic transcription. Given that such techniques have a useful role to play in a speech recognition system, it can be claimed that phonetic science should be able to contribute significantly to their development, both in their design and in the assessment of their performance. This paper deals with the latter application, discussing the extent to which automatic phonetic transcriptions can be accurately evaluated. This is discussed with reference to a system (which we call LUPINS) developed at Leeds University [7] which carries out speaker-independent Broad Class analysis of continuous speech by automatic segmentation and labelling; while the system was developed using a corpus of recordings from 18 speakers, the tests reported below were carried out with new data from new speakers and the system's recognition rules were left unaltered. It is claimed that a computational technique for measuring accuracy as outlined here will make testing much more efficient than a "manual" equivalent [8], and should be valuable in making explicit some of the phonetic principles underlying the analysis.

RECORDING OF ERRORS IN SEGMENTATION AND LABELLING

As explained in Roach et al (op cit), errors in transcription will be of a number of different types: (i) a segment is omitted; (ii) a spurious segment is inserted; (iii) a segment is assigned to the wrong phonetic category; (iv) a segment boundary is located incorrectly on the time axis. All of these errors must be detected and recorded in the assessment procedure, and some score reflecting the level of seriousness of the error must be derived. In our present research work (funded by S.E.R.C./Alvey Grant MMI-053) the assessment is carried out by a computer program which takes a transcription of a passage made by a human expert and compares it with the computer's transcription of the same data. The human transcription is always treated as the correct model (though it sometimes happens that the computer's version causes humans to revise their transcriptions). Since the transcription is typed in in the symbols of the Edinburgh Machine-readable Phonemic Alphabet or the "Alvey" ASCII symbol codes [8], while the computer transcribes using only a very small set of symbols (basically comprising Fricative, Nasal, Vowel, Dip, Stop, Flap, Burst, Silence), it is necessary for the human transcription to be converted into this alphabet before the comparison begins. All segments in both transcriptions are given duration values in csec.

A simple form of assessment was used in our earlier work: each error of types (i) to (iii) above was counted as one error, and a final success rate was arrived at by expressing the total number of errors as a percentage of the total number of segments in the passage. Errors of type (iv) were ignored. Scoring on this basis gave success rates in the region of 80% for informal conversational speech in six different languages with a number of different speakers including female and male. However, it was found that there were many cases where we felt we should treat some errors as "minor" or "forgiveable" (e.g. inserting a very brief Dip (approximant) segment between neighbouring voiced segments, or categorising a sound as a flap when the human had heard it as a brief stop), while other errors were considerably more serious; it was also found that the process of "marking" a machine transcription was a very time-consuming process that needed to be done after each run of LUPINS. It was because of these factors that it was decided to develop an automatic assessment technique. An additional advantage of doing this was that the technique should also make it possible to align an unknown recording of speech with its transcription: this has a number of potential applications in the field of large speech databases.

AUTOMATIC ASSESSMENT OF ACCURACY: EXAMPLES

Two short recorded test passages that were analysed recently are used as examples of the technique. The first passage is by two speakers, one male and one female, and the text is as follows:

M. Hello, operator - operator?

F. Yes, what can I do for you?

M. I'd like to make a telephone call.

The second passage is a male speaker saying "Can you recognise this sentence?". The assessment is done as follows:

(a) The human transcription (H) and the machine transcription (M) are compared symbol by symbol, and each case of matching symbols is scored as one correct symbol.

(b) When a symbol of H is found not to be matched by the corresponding symbol of M, the M transcription is corrected in one of the following ways:

(i) if M has missed a symbol, the symbol from H is inserted, and one error is recorded.

(ii) if M has inserted a symbol that is not pesent in H, that symbol is deleted and one error is recorded.

(iii) if the corresponding M symbol does not match, but subsequent pairs of H and M symbols do match, the M symbol is marked as incorrect, and is replaced by the H symbol. A score for the error between 0 (insignificant) and 1 (complete failure of recognition) is calculated by the procedure described in Section 4 below and added to the errors total.

(c) If the time values of the H segments are known (they are always included in transcription files made within our project, but may be missing from other transcriptions), the time values in M are adjusted to fit them, and the extent of the required adjustment is noted and added to a time-adjustment total score; adjustments in either direction on the time axis are treated as positive numbers. This score is kept separate from the scoring of correct/incorrect segments. Time measurement is done in csec, and the final time-adjustment error score is the number of csec recorded in the time-adjustment total as a percentage of the overall number of csec in the entire passage.

A particular case of a "missed symbol" is found fairly frequently when an intervocalic segment is missed and a very long vowel recorded instead. In the example given below, for example, the H sequence / əɹeɪ / should have been transcribed as VDV , but came out as a long V ; this would result in two errors being recorded, but we feel it is more appropriate to count this as a case of one missed segment.

MEASUREMENT OF ERROR GRAVITY

Our treatment of cases of incorrect symbols in the M transcription is still at a provisional stage, but it is clear that what is needed is some form of distance measure so that a wrong symbol that denotes a segment radically different from the correct one will be counted as nearer the error value 1, and a symbol that is not so different will receive a score that is nearer to zero. We measure distance by comparing segments on a feature by feature basis: in earlier work [8] we used phonetic features based on those of Ladefoged [10], but found difficulties in relating some of the features to our labels (which are essentially defined in acoustic terms) [11]. We are currently working with a set based on those used in the study of perceptual confusions among English consonants by Miller and Nicely [12]: the provisional set of five "primitive" features comprises +/- Voiced; +/- High energy (the term "high" is deliberately ambiguous between "high in amplitude" and "high in frequency", and is used to distinguish / s and ʃ / from other fricatives); +/- Nasal; +/- Transient (non-transient sounds are capable of having an audible steady state, while transients include plosives, bursts, semivowels and flaps) and +/- Fricative. The features could in some cases be given numerical (non-binary) values if wished, but for the purposes of this paper only binary values are used. (It is noticeable that even this small set contains more redundancy than phonologists would approve of). For each feature that was wrong in the M transcription, .2 was added to the overall error score, and the same was added to the "segments correct" total for each feature correctly identified: hence a case of all five features being wrong (e.g. Burst instead of Nasal) would cause 1 to be added to the total error score. On this basis, eight clear cases of error were selected for illustration and were scored as shown in Table 1, where the columns are headed 'H' for the human transcription using I.P.A. symbols, 'CME' for the "correct machine equivalent" (i.e. what the machine should have produced), 'WM' for the wrong machine transcription and 'S' for the error score for that segment.

TABLE 1
Examples of Error Scores

H	CME	WM	S
l	D	Fw	.6
d	S	Fm	.6
j	D	Fm	.8
d	S	D	.2
ʰ	B	Fs	.4
g	Fw	D	.6
Sil	Sil	S	.4
n	N	D	.6

RESULTS

Space does not allow a full presentation of the analysis of the example passages, but we will discuss one section: the first part is "Hello, operator" /heləu ɒpəreɪtə /, which in equivalent machine symbols is FVDVVSBVDVSBV . Table 2 shows the H transcription ('H'), converted machine equivalent symbols ('CME'), durations (D1), the actual machine-transcribed symbols ('M') and their durations (D2). The right-hand column gives our evaluation. The error score for the extract chosen is calculated as 4.2, with 9.8 correct symbols, giving a success rate of 57%. The time-alignment score is calculated

TABLE 2

Sample Assessment of Errors

H	CME	D1	M	D2	Result	ErrorScore
h	F	7	Fw	7	correct	
e	V	10	V	18	correct	
l	D	9	Fw	14	wrong	.6
əʊ	V	37	V	16	correct	
			D	12	spurious	1
ɒ	V	24	V	16	correct	
p	S	5	S	9	correct	
ʰ	B	2	B	2	correct	
ə	V	6	V	33	correct	
ɹ	D	7	-	-	missed	1
eɪ	V	16	(V)	(")	(continuation)	
t	S	6	Fm	6	wrong	.6
ʰ	B	3	-	-	missed	1
ə	V	28	V	31	correct	

as 59%.

Overall scores for the whole of the chosen test material were calculated on the same basis: 92 segments were processed, with a success rate of 60%. On time-alignment, a total of 954 csec of speech was processed, with a success rate of 72%.

It is clear from the figures that our automatic segment marking is stricter than our previous technique: this is probably not a serious matter, since our chief concern is to have a technique that is reliable and objective, and which allows us to make comparative judgments about system performance under different conditions. More work to refine the technique is, however, still needed.

REFERENCES

[1] D.A.Klatt, 'Overview of the ARPA speech understanding project', in W.A.Lea (ed.) Trends in Speech Recognition, Prentice Hall, 1980.

[2] D.W.Shipman and V.W.Zue, 'Properties of large lexicons', IEE-ICASSP, 1982, pp.546-549.

[3] R.A.Cole, R.M.Stern and M.J.Lasry, 'Performing fine phonetic distinctions: templates vs features', in J.S.Perkell and D.A.Klatt (eds.) Invariance and Variability in Speech Processes, Erlbaum, 1986.

[4] J.Vaissière, 'Speech recognition: a tutorial', in F.Fallside and W.A.Woods (eds.) Computer Speech Processing, Prentice Hall, 1985.

[5] W.Jassem and P. Domagala, 'Phonetic segmentation in a bottom-up automatic speech analysis',in Proceedings of the International Conference on Speech Input/Output, Institute of Electrical Engineers, 1986.

[6] J.Dalby, J.Laver and S.M.Hiller, 'Mid-class phonetic analysis for a continuous speech recognition system', in Proceedings of the Institute of Acoustics, 8.7, pp.347-354, 1986.

[7] H.N.Roach and P.J.Roach, 'Automatic identification of speech sounds from different languages', Working Papers in Linguistics & Phonetics, University of Leeds, 1983.

[8] P.J.Roach, H.N.Roach and A.M.Dew, 'Assessing accuracy in automatic identification of phonetic segments', in Proceedings of the International Conference on Speech Input/Output, Institute of Electrical Engineers, 1986.

[9] J.C.Wells, 'A standardized machine-readable phonetic notation', in Proceedings of the International Conference on Speech Input/Output, Institute of Electrical Engineers, 1986.

[10] P.Ladefoged, A Course in Phonetics,(2nd ed.), Harcourt Brace Jovanovich, 1982.

[11] P.J.Roach, 'Rethinking phonetic taxonomy', in Working Papers in Linguistics & Phonetics, vol.4, 1986, Leeds University (to appear in Transactions of the Philological Society,1987).

[12] G.A.Miller and P.E.Nicely, 'An analysis of perceptual confusions among some English consonants', J.Ac.S.,27.2, pp.338-352, 1955.

ETIQUETAGE AUTOMATIQUE DU SIGNAL DE PAROLE CONTINUE A L'AIDE DE LA VARIATION RELATIVE D'ENERGIE DES SEQUENCES DE PHONEMES

DESI M. RINGOT P. ANDREEWSKY A.

CNRS - LIMSI - ORSAY BP 30 91406 ORSAY CEDEX FRANCE

A threshold-free system for automatic labelling of speech signal is described. Mainly we transform the phonetic strings into energetic strings, using context-based rules or a square matrix which formalises the relative variation of energy between any two phonemes. For 700 sentences database, 95% of the labels are well matched. The adaptation, for other languages is easy.

L'étiquetage automatique, c'est à dire l'attribution de valeurs phonétiques aux spectres obtenus à partir d'un signal de parole, a pour but d'amorcer la phase d'apprentissage indispensable pour effectuer un décodage acoustico-phonétique performant pour la reconnaissance de la parole continue sur des vocabulaires étendus. Dans le présent travail on rappelle d'une part les idées générales de l'étiquetage automatique du système SHERPA et on expose une nouvelle version du module de calcul du profil théorique de la courbe d'énergie lorsque l'on connait la chaîne phonétique correspondante. Cette modification a comme intérêt d'une part de mieux exprimer la théorie sous jacente à l'étiquetage dans SHERPA et, d'autre part, de permetre une extension plus facile de cette approche à d'autres langues.

On donne les résultats de l'étiquetage automatique sur 700 phrases utilisant un vocabulaire de 2000 mots différents.

I. TRANSFORMATION DE LA CHAINE PHONETIQUE EN UNE SUITE D'ALTERNANCES PHONETIQUES, A L'AIDE DE REGLES CONTEXTUELLES.

Les opérations suivantes sont effectuées:

I.1. Définition des classes phonétiques.

Elle repose sur le principe suivant: deux phonèmes qui ont, dans un contexte phonétique identique, un comportement identique sur la courbe d'énergie, appartiennent à la même classe.

O (occlusives) : /p,t,k,b,d,g/
F (fricatives) : /f,v/
S (sifflantes) : /s,z/
X (chuintantes) : /ʃ,ʒ/
N (nasales) : /m, n, ñ/
L (liquides) : /l, r/
I (semi-voyelles) : /y, w, u/
V (voyelles) : /a, e, ε; i,ɔ̃
o, u, y, ø, oe, ɛ̃, ɑ̃, e muet/
DEB, FIN

I.2. La chaîne phonétique prononcée est transformée en une chaîne de classes.

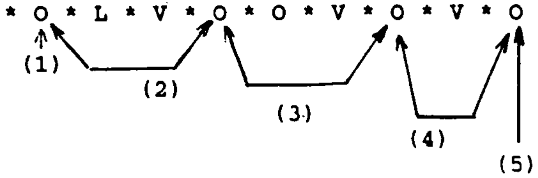
I.3. On écrit des règles phonétiques qui, à la chaîne de classes, va faire correspondre une suite de minimums (min), de maximums (Max) et d'alternances secondaires (Alts), à partir de laquelle on va calculer les paramètres de lissage de la courbe d'énergie, afin de transformer cette dernière en une suite de min, Max, Alts, ayant une interprétation phonétique.

I.4. Les règles phonétiques opèrent sur les classes et ont par exemple la forme suivante: (le symbole "*" signifie: suivi de)

1) Une occlusive en début d'énoncé prend la valeur 0.
2) Si O1*L*Y*O2
alors O1=0; L=½; V=1; O2=0
3) Si O1*O2*V1*O3
alors O1,O2=0 à 0; V1=1
4) Si O1*V1*O2 alors O1=0; V1=1
5) Une occlusive en fin d'énoncé prend la valeur 0.

Appliquées à la chaîne phonétique correspondant à "promptitude",
* p * r * ɔ̃ * p * t * i * t * y * d

* O * L * V * O * O * V * O * V * O



ces règles nous donnent la suite de 0,1,1/2,à suivante:

0 1/2 1 0 ⓐ 0 1 0 1 0

que nous appelerons "chaîne des contrastes". Cette chaîne fournit le nombre théorique d'alternances qu'il faut conserver sur la courbe d'énergie et par conséquent le nombre de fluctuations qu'il va falloir lisser sur cette courbe.

Ces règles contextuelles, au nombre de 2000 environ, attribuent des valeurs du type min aux occlusives, Max aux voyelles, min, Max ou Alts aux autres classes de phonèmes en fonction du contexte phonétique immédiat de gauche ou de droite. Les étiquettes attribuées peuvent dans certains cas regrouper plusieurs étiquettes simples (/yɔ̃/).

Les règles transforment toute chaîne phonétique, quelqu'en soit la longueur, en une suite d'alternances du type (0,1,0) qui correspondent à des pseudo-syllabes (dont la définition ne se superpose pas avec celle de la syllabe classique).

Comme on l'a vu dans l'exemple ci-dessus, les valeurs ½ et à peuvent s'insérer dans l'alternance (0,1,0), ce qui correspond à des fluctuations intrasyllabiques possibles. Le nombre d'alternances (0,1,0) et le nombre total de ½ et de ⓐ sont déterminants pour la procédure de lissage.

II. TRANSFORMATION DE LA CHAINE PHONETIQUE A L'AIDE D'UNE MATRICE DES CONTRASTES D'ENERGIE DES PHONEMES.

L'approche par règles contextuelles suppose une étude exhaustive des contextes phonétiques, assez longue à mettre en place pour une langue donnée et son adaptation ensuite à d'autres langues reste complexe. Nous exposons une méthode différente qui repose sur une meilleure définition théorique du problème, et qui est plus facilement généralisable.

Le principe de base consiste à utiliser l'évolution de l'énergie d'un phonème à l'autre, au cours de l'émission d'une chaîne phonétique continue.

Dans cette approche, les voyelles se situent toujours aux maximas d'énergie. Deux voyelles ou plus, qui se suivent sans hiatus, sont regroupées en un seul maximum. Tant que, à partir d'une voyelle ou d'un groupe de voyelle, l'énergie théorique des phonèmes successifs décroit, on est toujours dans le maximum. Par exemple, dans /artist/, le premier maximum est constitué par /ar/. Dès que l'énergie se met à croître, c'est que l'on est passé par un minimum (dans /artist/, on décroit à partir de /a/ en passant par /r/ pour aller jusqu'à /t/ et on remonte à partir du /t/ qui est un minimum énergétique). A partir du minimum d'énergie, et jusqu'au maximum vocalique suivant il peut s'insérer des alternances secondaires (½ ou ⓐ): occlusive suivie de liquide, occlusive suivie d'occlusive, occlusive suivie de n'importe quelle consonne.

Pour l'application de ces principes, on a utilisé une matrice carrée de dimension 10: 8 classes phonétiques et 2 symboles de début et fin d'énoncé.

Nous allons nous contenter de donner la partie de la matrice indispensable pour traiter la chaîne phonétique "promptitude" /prɔ̃ptityd/

	O	L	V	...	DEB	FIN
O	ⓐ	½	+	...	ø	ø
L	-	-	+	...	ø	ø
V	-	-	=	...	ø	-
...				...		
DEB	ø	ø	+	...	ø	ø
FIN	ø	ø	ø	ø	ø	ø

La signification de ces symboles est simple:

- Sur la première colonne (par laquelle on entre dans la matrice), on trouve la classe du phonème de gauche.
- Sur la première ligne, on trouve la classe du phonème de droite.
- "+" signifie que le phonème de droite a une énergie supèrieure à celle du phonème de gauche.

- "-" signifie que le phonème de droite a une énergie infèrieure à celle du phonème de gauche.
- "½" signifie que l'on peut avoir une alternance secondaire du type ½ entre phonème de gauche et phonème de droite.
- "=" signifie que les phonèmes de gauche et de droite ont une énergie similaire au sens de l'algorithme.
- "ø" correspond à une séquence impossible.

Si on utilise cette matrice pour transformer la chaîne phonétique /prɔ̃ptityd/, ou plus exactement sur la chaîne de classes phonétiques correspondante (O L V O O V O V O), on commence par attribuer la valeur 0 aux occlusives et la valeur 1 aux voyelles; ensuite on attribue à L dans O L V la valeur ½ et on écrit un à entre les deux occlusives de la séquence O O; on obtient finalement la séquence :

0 ½ 1 0 ⓐ 0 1 0 1 0

III. LE LISSAGE.

L'objectif du lissage est de localiser les spectres correspondants aux étiquettes des minimums (consonnes ou groupes consonnantiques) et aux étiquettes des maximums (voyelles ou groupes vocaliques). Les spectres correspondants aux fluctuations intrasyllabiques sont délibérément lissés mais ensuite, éventuellement, étiquetés après un réexamen de l'aspect de la courbe à l'intérieur de la pseudo-syllabe.

Pour cela on effectue un double lissage, dont l'intensité est guidée par la structure de la chaîne des contrastes, d'abord sur l'axe des énergies puis ensuite sur celui des temps. L'importance relative des lissages énergétique et temporel peut être paramètrée.

Remarque: il est indispensable d'effectuer un lissage sur les deux axes. En effet un lissage sur l'énergie seule risquerait de gommer les fluctuations liées à des pseudo-syllabes peu contrastées sur l'énergie (par exemple: /si/, /my/). Un lissage sur l'axe des temps, outre le fait qu'il pallie à l'inconvénient précédent, permet d'éliminer des fluctuations petites sur l'axe des temps mais parfois relativement importantes sur l'axe des énergies (par exemple l'explosion du /k/).

Le lissage s'effectue donc en deux temps et utilise deux nombres fournis par la chaîne des contrastes; le premier correspond au nombre total de pseudo-syllabes et donne le nombre d'extremums qu'il convient de conserver après les deux lissages; le deuxième correspond au nombre de fluctuations intrasyllabiques possibles et est déterminant pour évaluer l'importance relative des deux lissages. Cette importance relative est également régulée par un paramètre d'ajustement, si les critères de contrôle (Cf ci-dessous) de la bonne qualité de l'étiquetage ne sont pas vérifiés.

Le double lissage est obtenu par suppression itérative des fluctuations énergétiques puis temporelles jusqu'à obtention du nombre théorique d'extremums.

Soulignons que dans cette procédure de lissage, il n'est fait appel à aucun seuil ni sur l'énergie, ni sur le temps, ce qui représente un facteur de portabilité intra et multilocuteurs très important.

IV. PROCEDURE D'ETIQUETAGE.

La procédure d'étiquetage est relativement simple.

Dans un premier temps, on attribue aux extremums sélectionnés par le lissage, des étiquettes phonétiques consonnantiques (simples ou multiples) aux minimas et vocaliques (simples ou multiples) aux maximas. La valeur et l'ordre des différentes étiquettes sont donnés par la procédure de transformation des chaînes phonétiques.

Dans un deuxième temps, on essaye, à l'intèrieur même de la pseudo-syllabe, de dissocier les étiquettes multiples chaque fois que la présence d'une fluctuation énergétique intrasyllabique rend cela possible. Par exemple, l'étiquette complexe /pr/ dans la pseudo-syllabe /prɔ̃/, pourra être dissociée en /p/ et /r/ si entre les extremums correspondant au min: /pr/ et max: /ɔ̃/, il existe une fluctuation de la courbe. Il existe ainsi, compte tenu de l'énergie relative des phonèmes constituant des étiquettes multiples, plusieurs cas de figure que nous ne détaillerons pas ici.

V. PROCEDURES DE CONTROLE.

Ces procédures détectent au sens de certains critères, les phrases qui présentent un risque d'étiquetage défectueux. On propose alors une solution de réétiquetage en faisant varier le paramètre d'ajustement qui contrôle l'importance relative des lissages énergétiques et temporels.

La qualité du nouvel étiquetage est à son tour vérifiée sur l'ensemble des critères et il est remis éventuellement en cause. On effectue ainsi au plus six tentatives (le paramètre d'ajustement varie six fois); si à la sixième tentative les critères ne sont toujours pas vérifiés, la phrase est automatiquement rejetée du corpus d'apprentissage.

Actuellement seuls deux critères sont opérationnels. L'un vérifie que les N occlusives d'un énoncé sont placées sur les N minimas les plus bas de la courbe d'énergie; il détecte les erreurs globales d'étiquetage. L'autre vérifie que deux étiquettes consécutives sont séparées par un minimum de deux spectres; ce critère détecte la plupart des erreurs purement locales.

VI. RESULTATS.

L'étiquetage automatique a été testé sur un corpus de 700 phrases de longueur variable de 5 à 10 mots. La qualité de l'étiquetage est évaluée par rapport aux performances de l'étiquetage manuel par un phonéticien; les occurences d'un même mot à des parties différentes du corpus ont la même courbe d'énergie et les étiquettes correspondantes sont placées aux mêmes endroits de la courbe.

Phrases rejetées (critères de contrôle non vérifiés): 15%.
Etiquettes bien placées: 95%.
Etiquettes complexes: 15%.
complexes dissociées: 50%

Les 5% d'erreurs d'étiquetage ont peu de répercussion sur l'ensemble du système car les autres modules de l'apprentissage comportent des contrôles internes qui permettent de les détecter.

L'adaptation de ce système d'étiquetage à une autre langue que le français est simple. Il suffit de modifier le contenu de la matrice de transformation des chaînes phonétiques, en fonction du système phonétique de la langue. L'adaptation est en cours de réalisation pour l'espagnol et l'italien.

BIBLIOGRAPHIE

ANDREEWSKY A. DESI M. FLUHR C. POIRIER F. "Une méthode de mise en correspondance d'une chaîne phonétique et de sa forme accoustique", 11ème ICA, Revue d'Accoustique, 1983, p. 245.

ANDREEWSKY A. DESI M. POIRIER F. "Le système SHERPA -de l'étiquetage phonétique automatique à la reconnaissance par analyse ternaire", 5ème Congrès RFIA, 1985, p.

DESI M. POIRIER F. "Le système SHERPA: étiquetage et classification automatique par apprentissage pour le décodage automatique et la parole continue", Thèse de Doctorat en Sciences, Paris-Sud Orsay, 1985.

LENNIG N. "Automatic alignment of natural speech whith a corresponding transcription", Speech communication, 1983, p.190-192.

MERCIER G. "Accoustic-phonetic decoding and adaptation in continuous speech recognition", Automatic Speech Analysis and Recognition, Reidel Publishing Co, 1982.

WAGNER M. "Automatic labeling of continuous speech whith a Given Phonetic Transcription using Dynamic Programming Algorithms", IEEE Acoustics Speech and Signal Processing, Catalog N°81CH1610-5, 1981, p.1156-1159.

SEGMENTATION ET RECONNAISSANCE EN PAROLE CONTINUE A L'AIDE DES REFERENCES ISSUES DU SYSTEME VARAP.

RINGOT P. ANDREEWSKY M. DEVILLERS L. DESI M. PARISSE C.

CNRS - LIMSI - ORSAY BP 30 91406 ORSAY CEDEX FRANCE

We present two possible approaches of continuous speech recognition. The first uses a segmentation obtained by a training process. The second using an appropriate distance allows to simultaneously achieve the segmentation and recognition.

Dans ce travail, on expose les expériences de reconnaissance qui ont été faites, compte tenu du système d'étiquetage automatique employé (utilisé sur un corpus de 700 phrases) et du mode de sélection utilisé dans le système VARAP. Deux méthodologies différentes sont exposées. L'une qui procède d'abord à une segmentation, puis à une reconnaissance, la seconde effectue ces deux opérations simultanément. Dans ce qui suit, on utilise une distance qui est donnée par la formule :

$$|(O1 - X1) - (O2 - X2)| + \ldots + |(O15 - X15) - (O16 - X16)|$$

où les valeurs O1...O16 sont les 16 valeurs du premier spectre et X1...X16 sont les valeurs du second.

I. SEGMENTATION OBTENUE A PARTIR DU CORPUS D'APPRENTISSAGE.

Quatre paramètres sont définis: Em, Tm, EM, TM dont la signification est la suivante:

Em est un seuil minimal d'énergie.
Tm est un seuil minimal temporel.
EM est un seuil maximal d'énergie.
TM est un seuil maximal temporel.

Le mode d'utilisation de ces quatre paramètres est le suivant :
Toutes les alternances qui ont une différence d'énergie inférieure à Em sont lissées tant que leurs fluctuations sur le temps ne sont pas supérieures à TM. De même, toutes les alternances qui ont une fluctuation sur le temps inférieure à Tm sont lissées tant que les fluctuations sur l'énergie ne sont pas supérieures à EM. Les paramètres Em, EM, Tm, TM sont déterminés sur le corpus d'apprentissage. Pour cela, on sélectionne le plus petit écart non lissé sur l'énergie et sur le temps dans chaque phrase étiquetée du corpus et on calcule pour chacun de ces écarts le nombre de fois dans le corpus où il a été conservé ou lissé Le plus grand des plus petits écarts de chaque phrase du corpus donne les valeurs de EM et de TM.
La sélection des seuils inférieurs Em et Tm se fait en imposant un rapport aussi optimisé que possible entre le nombre de fois où Em, Tm ont été lissés et le nombre de fois où ils ont été conservés, étant entendu que l'optimisation est définie par le plus petit pourcentage possible d'erreurs conservées sur le corpus.
Les résultats de la segmentation sur 50 phrases après optimisation sont les suivants :
Dans les 50 phrases, il y a en tout 1032 segments.
Le lissage optimisé laisse un nombre total d'ajouts égal à 29 et un nombre total d'élisions égal à 28.
Parmi les ajouts, il y a 14 segments qui correspondent ou bien à des répétitions du même phonème (par exemple /sss/), ou bien à des dissociations d'étiquettes complexes du type voyelle-voyelle (par exemple /aa/ qui donne /a/,/a/) ou bien des étiquettes complexes du type voyelle-liquide (par exemple /ar/ qui donne /a/,/r/).

Il reste donc 15 ajouts qui ne peuvent pas être interprétés dans le cadre du niveau phonétique où nous nous situons.
Pour les élisions, 18 segments correspondent à des étiquettes complexes que la segmentation n'a pas dissociées. Ce sont des groupes du type consonne-liquide ou consonne-chuintante (par exemple /b/ et /r/ formant une seule étiquette /br/) ou encore des groupes du type consonne-consonne (par exemple /t/ et /d/ ne formant qu'une seule étiquette /td/).
Il reste donc 10 segments d'élisions.
Par conséquent, le pourcentage total d'erreurs est de l'ordre de 5% . Si on se refère à une ségmentation phonémique, le pourcentage d'erreurs est de 3% dans notre système de référence qui admet des étiquettes phonétiques multiples.

II. UNE EXPERIENCE DE SEGMENTATION ET DE RECONNAISSANCE SIMULTANEE.

Cette expérience comprend les étapes suivantes:

A. Sur la suite continue des spectres correspondants à un énoncé donné, on prélève les triplets successifs de spectres en commençant dans l'ordre par le premier spectre puis le second et ainsi de suite. Deux triplets consécutifs ont donc toujours deux spectres en commun.

B. Ces triplets sont proposés au dictionnaire des références ternaires obtenues à partir du système VARAP. On obtient ainsi des treillis de quatre phonèmes candidats, résultats d'un scrutin majoritaire effectué sur les quinze plus proches références du dictionnaire où l'on tient compte de la position des références et de leur nombre.
Simultanément, on conserve la distance entre le meilleur candidat du treillis et la référence analysée, et on affiche la courbe des distances.

C. Résultats.
Un dépouillement effectué sur 200 phrases montre tout d'abord que la courbe des distances suit les contrastes de la courbe d'énergie et qu'elle fournit une segmentation de même qualité.

Cela nous conduit à faire la remarque importante suivante: l'analyse centiseconde effectuée utilise un dictionnaire de références qui ne contient pas les transitions. Par conséquent, on pouvait s'attendre à ce que les extrema de la coube d'énergie soient aux minima de la courbe des distances. Une explication de ce phénomène tient dans le fait que la distance utilisée bien qu'étant du type convergence uniforme contient un facteur d'énergie qui se manifeste de manière importante sur les voyelles.

Les comparaisons avec le corpus d'apprentissage font apparaître:

- des plages de grande stabilité phonétique permettant de déterminer des îlots de confiance.
- la possibilité en cours d'élaboration d'identifier l'énoncé à partir de la suite des treillis et de la disposition des extrema de la courbe des distances à l'exception des débuts et des fins d'énoncé.

Des dépouillements effectués sur 200 phrases donnent des résultats de reconnaissance phonétique très variables selon les phrases, de l'ordre de 60 à 75 % .
Les procédures indispensables permettant de dégager une décision à partir des treillis et des extrema sont en cours d'élaboration.

BIBLIOGRAPHIE

DESI M. POIRIER F. "Le système SHERPA: étiquetage et classification automatique par apprentissage pour le décodage automatique de la parole continue", Thèse en Sciences, Paris-Sud, Orsay, 1985.

LAZREK M. HATON J.P. "Segmentation et identification des phonèmes dans un sysème de reconnaissance automatique de la parole continue", Congrès AFCET Reconnaissance des Formes et Intelligence Artificielle, Paris, Janvier 1984, p. 5.

MARIANI J. "ESOPE: un système de compréhension de la parole continue", Thèse d'Etat, Université Paris VI, 9 juillet 1982.

MARIANI J. "Méthodes en reconnaissance phonétique", 11ème ICA, Toulouse, 125-137, 1983.

MERCIER G. GERARD M. GILLET D. NOUHEN-BELLEC A. QUINTON P. SIROUX J. "Le système de reconnaissance de la parole continue KEAL", 12ème JEP, GALF, Montréal, 1981.

MERCIER G. "Analyse acoustique et phonétique dans le système KEAL", 12ème JEP, GALF, Montréal, 1981.

THE ALGORITHM FOR THE PHONEMIC LABELLING AND SEGMENTATION OF SPEECH WAVEFORMS USING FEATURE MAPS

VLADIMIR CHUCHUPAL

Computer Center of the Academy of Sciences of the USSR
Moscow, USSR, 117967

ABSTRACT

In this work the algorithm for the phonemic labelling and segmentation of speech waveforms is described. This algorithm is founded on the feature maps: the self-organized neural networks model. The model is able to form automatically a representation of distribution of speech signal parameters. The algorithm described below utilizes this ability in order to form criteria of phonemic labelling and segmentation. In the such manner we produce the representation for not only short-time signal parameters but also of the temporary trajectories of this parameters.

INTRODUCTION

One of the most succesful speech recognition methods is one, which founded on the use of statistical laws,which has been established in the speech signal parameters distribution. Therefore, it seems important to investigate methods which is able to accomulate the data about distribution of speech signal parameters, for example, to approximate the probability density function of mutial distribution of this parameters. Often this task can be solved satisfactory by means of self-organizing neural network models, in particulary, the model for the self-organized formation of structured feature maps /1/.

Let us $\mathcal{L}$ be a pattern space, the elements of $\mathcal{L}$ may be represented by vectors $\bar{x} \in R$ (pattern vector). The structured representation of is formed with the help of matrix $M_{L \times L}$ (feature map) with the elements $\bar{m}_{ij}$. Every m_{ij} is defined by it's time-variable weights $\bar{m}_{ij} = (\mu^k_{ij})_{k=1,\dots,L}$. Initially, the values of the $\bar{m}_{ij}$ choosed in the randomly manner. An algorithm creation of features map consist of two steps /1/. Let us, for the time moment t, t=0,1, ..., n, ... the input pattern vector would be x(t). Then, in the first step, we define the indexes i_0, j_0 of the element $m_{i_0 j_0} \in M$, such, that:

$$\| \bar{x}(t) - m_{i_0 j_0} \| = \min_{i,j} \| \bar{x}(t) - m_{ij} \| \quad (1)$$

In the second step the modifications of weights μ^k_{ij} is made. For $m_{i_0 j_0}$ and its neighbours (for example, if the radius r(t)=1, the neighbours for the $m_{i_0 j_0}$ will be $m_{i_0+1\, j_0}, m_{i_0-1\, j_0}, m_{i_0\, j_0-1}, m_{i_0\, j_0+1}$):

$$\bar{\mu}_{ij}(t) = \bar{\mu}_{ij}(t-1) + \alpha(t)(\bar{x}(t) - \bar{\mu}_{ij}(t-1)) \quad (2)$$

In equation (2) $\alpha(t)$ satisfy the conditions: $\sum_{t=0}^{\infty} \alpha(t) = +\infty$; $\sum_{t=0}^{\infty} \alpha^2(t) < +\infty$; $\alpha(t) > 0$

It was shown /1,2/ that for correct choise the values of the $\alpha(t)$ and r(t), described above process has the next properties. When t $\longrightarrow \infty$ the values of μ^k_{ij} change so, that adjacent elemehts of the matrix M respond to (in the sense of equation (2)) closed (in the sense of norm $\| \cdot \|$) vectors from space $\mathcal{L}$. The distribution of values μ^k_{ij} on the matrix M approximates the mutial distribution probability density function for patterns vectors.

The successful application of feature maps for fonemic labelling have been made in the work /3/. But the fonemic qualities of the sounds depend not only of it's short-time spectra, but also the context - phonemic qualities of the adjacent phonemes. In our investigation the method , was described, and feature maps, produced in the such manner, was used for creation the segments boundary criteria and accomulation the information about temporary traectories of spectral parameters. It's apparently, that this information may be useful for transeme segments analysis.

AN AUTOMATIC FORMATIONS THE CRITERIA FOR SETTING THE LABELS OF THE SEGMENTS BOUNDARY IN THE SPEECH SIGNAL

We assume, that the important role in the speech perception belongs to the stationary segments of speech and the silence segments. This segments may be viewed the adaptation's signals for our hearing system in the sense of adaptation to amplitude spectra of the sound. Therefore, the labels setting, in order to mark the stationary segments, may be useful on one hand, to produce the phonemic identification this segments, and on the other hand, to correctly identify the transition segments, which phoneme interpretation depend on long-time information. As the input patterns we used the short - time spectra $S(\omega,t)$ and the phonemic function /5/: $\Phi(\omega,t) = lg(|S(\omega,t)| - |S(\omega,t-\tau)|)$, where ω denote frequency, t - time, and τ - small time delay. We use the FFT algorithm in order to calculate the 252 - point amplitude spectra (divided into 21 frequency channel in the range 40 Hz-5 kHz) every 12.6 ms. Central frequency of each channel was equally spaced and the channel 22 contented the total energy of the segment. The values of fonemic function calculated from two adjacent short - time spectra. We used the synthetic sounds. Three sounds modelled the vowels. This formant frequency were spaced at 900 Hz, 1600 Hz, 2900 Hz. One sound was represented as an unvoiced fricative.

On the first step we formed two maps: map for short - time spectra and the second map for fonemic function values. The matrix M contained 6 6 elements in both cases. The process (1)-(2) contained T=20000 steps. The values and decreased linearly: $\alpha(t) = \alpha_0(1 - t/T)$, $r(t) = r_0(1 - t/T)$ where $\alpha_0 =$, $r_0 =$. We denoted sounds stimulus as A, B, C, D. The resulting maps are shown in the figures 1 and 2. In order to denote the elements of maps the next procedure was applied/1/. Approximately one hungreed of well known patterns of every sound were presented to input the algorithm (1)-(2). The element, mainly corresponded (in accordance with (1)) to patterns of the sound A was denoted A.

C	-	-	D	D	D
-	-	-	C	-	A
A	-	-	A	A	A
A	-	A	-	-	B
-	-	-	-	-	-
A	-	-	B	-	B

Figure 1. The feature map for the values of the short - time spectra. The symbol '-' denotes the nolabelled elements.

N	-	-	-	-	-
-	-	-	-	-	-
N	-	N	-	-	-
N	N	N	-	-	-
-	-	N	-	N	-
S	-	N	-	-	-

Figure 2. The feature map for the fonemic function values. The symbol 'S' corresponds the values of fonemic functions for stationary segments of the sounds A,B,C. The symbol 'N' denotes the same segments of the sound D. The symbol '-' denotes the transition segments.

In the segmentation and labelling algorithm we supposed, that elements denoted by the symbols 'N' and '-' would be correspond to nonstationary segments.

In the work /6/ it have been made suggestion about existance of special cells - detectors for phonemes boundary detection. The first question was: may the map of fonemic function values to be use as as the map of such detectors ? We tested this capability of the map using the continious signal, contained 140 above - mentioned sounds. On the map of fonemic function values we obtained the trajectory, consisted of the elements, that corresponded the sequence of input patterns. The algorithm produced the label of transition region (segments) when this corresponded element was belong to transition region of the map (or, another words, was denoted as transition element). The labels of stationary regions were produced in the such manner. In the case of stationary regions the algorithm made an attempt to interpret this segments in accordance with the map of short - time spectra. The analysis of the result shows that all stationary segments belonged vowels sounds have been labelled correctly. About 8% of the transition regions were omissed. All stationary segments were recognized (with respect to map of short - time spectra) right.

THE USE OF THE INFORMATION ABOUT TEMPORARY TRAJECTORIES OF THE PARAMETERS FOR THE FONEMIC LABELLING

In order to use the temporary trajectories of the parameters as a feature patterns, let us see the next feature map (denote it map III) formation process. The input vectors for this map consisted from the values of the outputs of the map of short - time spectra (map I). The dimensionality of the input vectors to map III is equal to the number of elements in the map I, and the values of the components of this elements are equal to output values (see equation (1)) of the correspond elements of the map I (these output values have been added during some times). It can be said, that each element of the map III is connected with each element of map I. In order to control the map III formation process we used the map of phonemic function values (map II). When the corresponded element of the mapII was the element, denoted as ststionary, the label of stationary segment was produced. Up to this moment the values have been summing up and the result was used as the input vector to map III. The produced label element of map II became nonactive for some time. For the formation of the map III we used the map I and map II, described above. The number of elements in the map III was 4x4. The process contains T=6000 steps. The values of the radius r(t) and parameter $\alpha(t)$ where choosen as it was shown below. The result is presented on the figure 3.

BB	—	—	AA
—	—	—	AA
BC	AB	—	AA
CC	—	—	AB

Figure 3. The feature map for temporary trajectories of parameters. Here AA, BB, CC, are corresponding to the stationary segments, BC and AB are corresponding to transition regions.

In the test signal for formation of the map III we used transition region beetween A and B, B and C, C and D, D and A only. It is clear, from the figure 3, that no exist elements, that correspond to transition regions CD, DA and sound D. We tried to label the test signal, described above, with the help of the map III. In this case the algorithm was the same as the algorithm for the creation of the map III. The only difference between then was that in the algorithm of the labelling, every input vectors was identified in according with map III. As it was expected, we received 100 of correct detection of transitions between B and C, A and B and stationary segments of A,B,C. But the detection of the sound D and transition regions CD and DA contained many mistakes.

CONCLUSION

It have been shown in our works, that use of model of the feature maps formation yields the possibility to form in the simple manner the labelling and segmentation rules founded on statistical properties of the signal. This rule uses the properties both stationary and transition segments of signal.

REFERENCE

1. Kohonen T. Self - Organization and Associative Memory, Springer, 1983
2. Cottrell M., Fort J.C. A Stochastic Model of Retinotopy: A Self-Organizing Process. Biol.Cybern.,Vol.53, No 6, 1986.
3. Kohonen T., etc. Phonotopic maps - Insightful Representation of Phonological Features for Speech Recognition, Proc. of PRIP-84, Montreal, pp. 182-185, 1984.
4. Бондарко Л.В. Фонетическое описание языка и фонологическое описание речи. Л., Изд-во Ленинградского ун-та,1981
5. Пирогов А.А. К вопросу о фонетическом кодировании речи. Электросвязь, 1967, №5, с. 24-31.
6. Чистович Л.А., Венцов А.В., Люблинская В.В. Слуховые уровни восприятия речи. Функциональное моделирование. В сб.: Акустика речи и слуха. Л.,"Наука", 1986.

A LARGE BULGARIAN CENTRAL ALLOPHONES DATA BASE

PHILIP CHRISTOV

Voice Man-Machine Communication Lab.
Mechanics & Biomechanics Institute
Sofia, Bulgaria, 1090, P. O. Box 373

ABSTRACT

In this paper a large data base of Bulgarian central allophones is considered consisting of about 5000 utterings from 60 professional speakers (30 male and 30 female). They are imbeded in words each pronounced at the end of a standard carrier sentence. Professional analog (motion picture magnetic and optical) recordings are available of the carrier sentences together with the digital recordings (IBM compatible) of the vowel and consonant segments, manually extracted from the carrier. Each such record is labeled with a digital code which allows computer selecting, sorting and merging the data, according to the particular research purpose. The data base is verifyed by a group of 20 listeners by means of a semiautomatic aural identification procedure (CHRISTOV, Ph., Acustica, 29, 347-349, 1973).

INTRODUCTION

Contemporary computing technology offers vast prospects in processing large amounts of speech material in sensible stretches of time. The results are free of individual interpretations of the experimental data (as is the case, for example, by visual reading of spectrograms). If by the preparation of the machine speech input suitable knowledge is used, the machine output should be comparable not only within the limits of a given language, but also within the much broader limits of its language group.

SPEECH INPUT

A principal requirement in performing meaningfull acoustic measurements upon the phonetic units of speech /1/ is to use as experimental stuf PHONES (See REMARK) of comparable ALLOPHONES imbeded in words, uttered in equal phrases with equal intonation.

In agreement with this requirement the composite parts of the data base are phones of the central allophones of the Bulgarian speech, namely:

/C-V̀-C/ - stressed vowels in relevant environment of two bilabial consonants

/a-C-à/ - consonants in relevant environment of a preceeding unstressed /a/ and a following stressed /à/ /2/.

Each phone, together with its relevant environment is put into a major stress fraction of a word, taken from Bulgarian unilingual dictionaries and inserted at the end of a carrier sentence.

The carrier sentence consists of four phrases: The first phrase carries the label of the individual speaker, the second-the label of the phonemic group, the third-the label of the particular allophone and the last-the carrier word.

As shown in the TABLE except of the six Bulgarian vowels from the central allophones /bV̀b/ and /pV̀p/, there are vowels from two more allophones which are not central. The first of them, /tV̀t/ is chosen because common words could be found for all Bulgarian vowels. The second, /bVb/, is the unstressed opposition of the central allophone /bV̀b/. All consonant phones of the central allophones /aCà/ are contained in the same environment speech fraction /naCà/ which is imbeded in the initial part of the carrier word.

The consonant phones in initial and final position in words are presented by the allophones /Càn.../ and /...nàC/. The latter may as well be considered as central for the consonants of these particular phonetic groups. Judjing from the similar phonetic context there is little doubt about their comparability with the central allophones /naCà/.

BUILD UP PROCEDURE

The data base build up procedure is similar to that employed in the classical study of Potter and Steinberg /3/ the difference being in its intensification by application of computing technology (Fig.1).

The test utterings are extracted from two groups of speakers:

30 bass-bariton males
30 mezzo-soprano females

They all are professionals with distinct pronunciation selected among the radio and television speakers and the actors from the theaters of the Bulgarian capital. Their voices are energetic and belong to people between 25 and 50 years old. From them has been demanded to pronounce each phrase of the carrier sentence indifferently, in slow style and with falling intonation.

The working language in front of the microphone was Standard Bulgarian, i.e., the language the speakers are practizing during their public performances.

The speakers read the test material two times from randomly mixed cards to avoid the practice effect.

AUDITORY VERIFICATION

The auditory verification of the data base /4/ is carried on by a group of 20 lay listeners all native Bulgarians from different parts of the country and with some technological education.

The listeners have been presented two times with sound recordings of the carrier sentences. Having at their disposal listeners cards, containing labeled carrier words, they reacted by filling in the empty spaces in the carrier words with the letters of the phones they heard.

The listeners output was punched on machine cards and processed by a computer program which excludes the responses of the false listeners and punches on cards the labels of uncorrectly pronounced phones.

ANALOGUE DATA

During the sound recording session the speakers read the input cards with moderate voice effort and kept a sound level between 60 and 80 dB. The microphone was placed in the middle of a highly damped camera for acoustic measurements.

The audio recording was carried on by a professional sound recording staff which used studio equipment. The frequency response of the sound recording equipment via microphone and magnetic tape was flat between 100 and 15000 Hz (± 2 dB) and the noise level via magnetic tape was -64 dB. During the recordings the voltage and the frequency of the altrnating current network remained inside their standard limits: 220 V (+4.5%, -0%) and 50 Hz (±0%).

The original speech data are recorded on 6.35 mm magnetic tape with tape speed of 760 mm/sec. Working copies have been prepared on standard 35mm motion picture perforated magnetic and photographic tapes including high quality oscillograms and trivial variable-area photographic motion picture sound recordings. They can be displayed simultaneously and synchronously on motion picture sound editting equipment thus offering the technological prerequisite for simultaneous audio-visual inspection of the analog sound recordings.

SEGMENTATION

The "segmentation" of the analog data was carried on manually on a sound reading bench after carreful audio-visual inspection of the sound records and their control oscillograms.
It consists of marking the beginning and the end of each magnetic tape segment, carrying a labeled phone, with a perforation and a strong magnetic pulse (Fig. 2).
The places of the markers belonging to each such magnetic tape segment have been determined after:

1. The content of the segment has been HEARD as coextensive with the labeled-phone quality
2. The visual duplicate of the same segment has been OBSERVED on the control oscillogram or motion picture sound record:
   a) By the VOWEL-phones: As a mighty tone burst between two weak noisy signals
   b) By the CONSONANT-phones: As a moderate or weak noise-like signal between two mighty tone bursts or between a pause and a mighty tone burst.

DIGITAL DATA

The segmented analogue data were fed to the input of an analog to digital converter set "on" by each "Start" marker and "off" by the imediately following "Stop" marker. In the analog to digital converter the nonturbulent segments (vowels and resonants) were sampled with a rate of 20kHz and the turbulent ones with 40kHz to ensure no lose of audio frequency information below 10kHz, for the sonant-like, and below 20kHz for the turbulent sounds.
The output digital data files were stored in an IBM-compatible magnetic tape memory with recording density of 1600 bit/inch.
After the analog to digital conversion the data were processed by a servise program which does three things:

1. Records meaningfull 9-symbol labels at the head of the first block of each file. The simbols are decimal numbers carrying information about the origin, the hystory, the kind and the inventory of the file. The last symbol in the label is a key which lockes the file if ordered.
2. Performs correction of the segmentation.
3. Lockes unfinished files at the end of the tape or files for which no agreement was reached between the speaker and the group of reliable listeners.

CONCLUSION

The data base here considered was created for a research aimed at the build up of the phonetic fulcrum of the Bulgarian language for the purposes of computer recognition of continuous speech. It is available in 14 volumes, each stored on a 1/2 inch, 2600 feet magnetic tape together with a subroutine in FORTRAN IV for sequential or selected reading and/or rewriting of the labeled files. Its output can be easily reshaped by the use of standard sort/merge programs because in the labels are present explicite symbols except for the phonemic category but also for each individual speaker and its sex. This way the data base can be used for research purposes not only in the field of trivial and comparative acoustic phonetics but also in studying the prosodic features in speech connected with the personality and the sex of the speakers.

REMARK

From the different definitons of PHONEMES existing in the linguistic litterature the definition of B. Bloch /5/ is adopted here because it suits best the technological aspects of speech:

PHONEME - class of allophones
ALLOPHONE - class of phones in the same relevant environment
PHONE - any continuous fraction of a phrase that is heard as coextensive with a given quality
PHRASE - an utterance or part of an utterance bounded by successive pauses

REFERENCES

/1/ Twaddell, W.F., "Phonemes and Allophones in Speech Analysis", JASA,24,607-611(1952)
/2/ Stoikov, S., "Introduction into the Bulgarian Language Phonetics", 3-th rev.ed., Nauka i izkustvo,Sofia (1966) (In Bulgarian)
/3/ Potter, K.R., Steinberg, J.C., "Towards the Specification of Speech",JASA, 22,807-820(1950)
/4/ Christov, Ph., "A Semiautomatic Speech Sounds Aural Identification Procedure with Its Application to Speech Analysis", Acustica,29,347-349 (1973)
/5/ Bloch, B., "Studies in Colloquial Japanese", Language,26,86 (1950)

TABLE of involved allophones:

Test uttering position in the carrier word	Initial	Medial	Final
Stressed vowels	/bV̀b/ /pV̀p/ /tV̀t/	/ / /	/ / /
Unstressed vowels	/bVb/	/	/
Resonants	/Càn/	/naCà/	/nàC/
Voiced turbulent	/Càn/	/naCà/	/
Voiceless turbul.	/Càn/	/naCà/	/nàC/

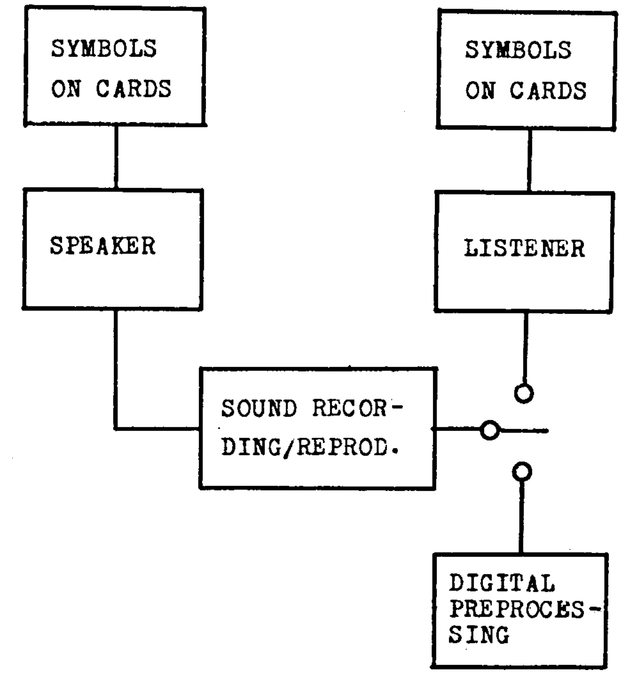


Fig. 1. Flow chart of the data base build up procedure

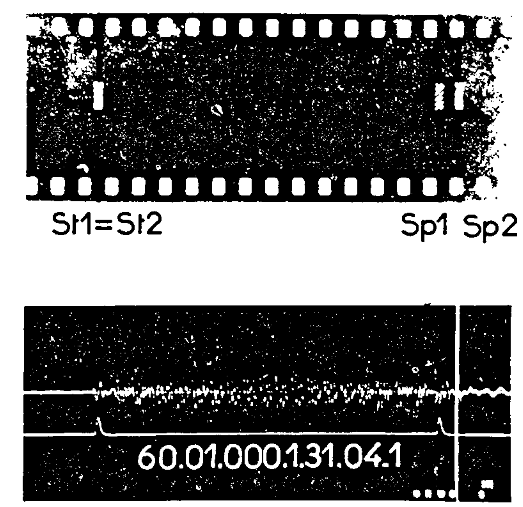


Fig. 2. (Top) Segment of magnetic tape record carrying the vowel /à/ from the allophone /bàb/ imbeded in the word /bàba/ uttered by speaker N.D.(male); (Bottom) Control oscillogram of the same segment
St1, Sp1 markers of primary segmentation
St2, Sp2 corrected segmentation

ФОРМИРОВАНИЕ БАНКА АПРИОРНЫХ ДАННЫХ О РЕЧИ ДИКТОРА

ВЛАДИМИР САННИКОВ   ЮРИЙ ПРОХОРОВ   ЮРИЙ ЖУРАВСКИЙ

Кафедра теории передачи сигналов, Московский электротехнический институт связи, Москва, СССР III024

АННОТАЦИЯ

Рассмотрена нелинейно-параметрическая модель речевого сигнала (РС). Дана методика оценки априорных данных, необходимых для оптимальной фильтрации РС. Приведены результаты исследования влияния априорных данных на качество фильтрации, а также за зависимости их от длительности обучающей выборки, смыслового содержания и способа произнесения речи диктора.

МОДЕЛЬ СИГНАЛА И АПРИОРНЫЕ ДАННЫЕ

Современное состояние теории и техники цифровой обработки и передачи речи характеризуется широким использованием марковских моделей РС [I,2]. На основе концепции переменных состояния динамической системы, К которой отнесем систему речеобразования, запишем марковскую модель РС в виде

$$\begin{aligned} x_{t+1} &= F(a_t)\cdot x_t + G\cdot y_t + B_x\cdot w_{x,t}, \\ a_{t+1} &= \Lambda_a\cdot a_t + \lambda_o + B_a\cdot w_{a,t}, \\ y_{t+1} &= F(c_t)\cdot y_t + \Psi(y_t) + B_y\cdot w_{y,t} \\ c_{t+1} &= \Lambda_c\cdot c_t. \end{aligned} \qquad (I)$$

Первое уравнение отображает голосовой тракт, второе – управляющие процессы или сообщения, характеризующие изменение во времени параметров голосового тракта; третье – источник квазипериодического возбуждения тракта; четвертое – параметры источника квазипериодического возбуждения.

В (I) t – дискретное время, x_t – вектор отсчетов РС, a_t – вектор отсчетов сообщения, y_t – вектор отсчетов квазипериодического сигнала возбуждения, c_t – вектор параметров источника квазипериодического возбуждения, $F(a_t)$ – квадратная матрица с элементами $F_{i,i+1} = I$, $i = \overline{I, m-I}$, $F_{m,i} = a_t^{(i)}$, $i = \overline{I,m}$, остальные её элементы равны 0, m – размерность векторов x_t и a_t; квадратная матрица $F(c_t)$ строится аналогично; Λ_a – квадратная матрица, характеризующая корреляционные взаимосвязи марковских сообщений; λ_o – постоянный вектор, который вместе с матрицей Λ_a характеризует среднее значение a_t; $\Psi(y_t)$ – нелинейная функция, обеспечивающая квазипериодический характер сигнала y_t [2]; G, B_x, B_a, B_y, Λ_c – постоянные матрицы; $w_{x,t}$, $w_{a,t}$, $w_{y,t}$ – случайные некоррелированные векторы гауссовских шумов с нулевыми средними и единичными матрицами корреляций. Для полного статистического описания поведения системы речеобразования необходимо задать начальные значения векторов средних значений и корреляционных матриц переменных состояния.

Объединяя переменные x_t, a_t, y_t, c_t в один блочный вектор состояния

$$z_t^T = (\, x_t^T;\ a_t^T;\ y_t^T;\ c_t^T \,)$$

запишем уравнение наблюдения в виде

$$u_t = h(z_t) + v_t, \qquad (2)$$

где u_t – наблюдаемая последовательность, $h(z_t)$ – скалярная функция векторного аргумента, v_t – случайная последовательность гауссовского шума наблюдений с нулевым средним и заданной дисперсией.

Соотношения (I) и (2) позволяют применить алгоритмы марковской фильтрации для выделения x_t из u_t с одновременным оцениванием процессов a_t, y_t, c_t. При этом синтез оптимальных алгоритмов фильтрации возможен только при известных характеристиках РС и его параметров. К таким априорным данным о речи диктора относятся матрицы: G, B_x, Λ_a, λ_o, B_a, B_y, Λ_c.

Отсутствие достаточно полных сведений о структуре и параметрах системы управления движением артикуляционного аппарата не позволяет воспользоваться моделью (I), так как указанные матрицы заранее не известны. Поэтому встает задача экспериментального их определения.

В качестве приближения к указанным априорным данным предлагается использовать их локально-постоянные оценки, которые можно получить при обработке фрагмента записи РС заданного источника в отсутствии помех. Совокупность отсчетов фрагмента РС образует обучающую выборку. Собственно процедура оценки элементов указанных выше матриц и называется формированием банка априорных данных о речи диктора.

МЕТОДИКА ОЦЕНКИ АПРИОРНЫХ ДАННЫХ

Фрагмент незашумленного РС разбивается на сегменты, каждый длительностью $T_c = I0 \div 20$ мс. Предполагается, что оцениваемые величины постоянны на сегменте анализа. Каждый сегмент обучающей выборки классифицируется по признаку "тон-шум". На участках типа "тон" оцениваются средние периоды основного тона T_0 и первой форманты T_I. Если c_t двумерный вектор, то его составляющие $c_t^{(1)}$ и $c_t^{(2)}$ можно выбрать так, чтобы при заданных вектор-функции $\Psi(y)$ и коэффициентах B_Ψ и B_y дискретное векторное колебание $\hat{y}_t$ имело собственную частоту $\hat{F}_1 = I/T_I$ на периоде $\hat{T}_0 = T_0$. Далее по оценкам $\hat{c}_t$ оцениваются элементы λ_{cij} матрицы $\hat{\Lambda}_c$.

Оценка вектора коэффициентов $\hat{a}_t$ производится по обучающей выборке на основе известных алгоритмов идентификации параметров стохастических систем [3,4]. Траектории оценок $\hat{a}_t$ являются тем материа –

лом, который необходим для вычисления оценок $\hat{\Lambda}_a$, $\hat{\lambda}_o$, $\hat{B}_a$. С этой целью по траекториям $\hat{a}_t$ вычисляются векторы средних $\bar{\hat{a}}$ и стандартных отклонений $\hat{\sigma}_a$. Переходя к центрированным и нормированным траекториям $\tilde{a}_t^{(i)}=(\hat{a}_t^{(i)}-\bar{\hat{a}}^{(i)})/\hat{\sigma}_a^{(i)}$, $i=\overline{1,m}$, полагаем, что они удовлетворяют стохастическому уравнению

$$\tilde{a}_{t+1}=\tilde{\Lambda}_a\cdot\tilde{a}_t+\tilde{B}_a\cdot w_{a,t}. \qquad (3)$$

Полагая $\tilde{a}_t$ эргодическими процессами, определим усреднением по времени взаимнокорреляционные матрицы

$$E\tilde{a}_{t+1}\cdot\tilde{a}_{t+1}^T=\tilde{\Lambda}_a\cdot R_a^{(1)}+\tilde{B}_a\cdot\tilde{B}_a^T=R_a^{(0)},$$
$$E\tilde{a}_t\cdot\tilde{a}_{t+1}^T=\tilde{\Lambda}_a\cdot R_a^{(0)}=R_a^{(1)}, \qquad (4)$$

откуда находим

$$\tilde{\Lambda}_a=R_a^{(1)}\cdot[R_a^{(0)}]^{-1},$$
$$\tilde{B}_a\cdot\tilde{B}_a^T=R_a^{(0)}+R_a^{(1)}\cdot[R_a^{(0)}]^{-1}\cdot R_a^{(1)}. \qquad (5)$$

На основе (3) и (5) с учетом $\bar{\hat{a}}$ и $\hat{\sigma}_a$ получаем искомые оценки матриц в виде

$$\hat{\Lambda}_a=D_\sigma\cdot\tilde{\Lambda}_a\cdot D_\sigma^{-1},$$
$$\hat{\lambda}_o=\bar{\hat{a}}\cdot(1-\hat{\Lambda}_a);\quad \hat{B}_a=D_\sigma\cdot\tilde{B}_a, \qquad (6)$$

где $D_\sigma=diag[\hat{\sigma}_a^{(1)}\ldots\hat{\sigma}_a^{(m)}]$ – диагональная матрица стандартных отклонений.

Важным вопросом методики оценки априорных данных является расчет необходимой длительности обучающей выборки. Она равна $T_{ов}=M\cdot T_c$, где M – минимально необходимое число сегментов на фрагменте анализа. Установлено, что $M\approx 15\cdot\Delta_g^{-2}$, где Δ_g – доверительный интервал для элементов матриц $R_a^{(0)}$ и $R_a^{(1)}$. При $\Delta_g=0,1$, $T_{ов}=10\div20$ с. Дальнейшее увеличение $T_{ов}$ не приводит к значительному изменению оценок данных.

АНАЛИЗ ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ АПРИОРНЫХ ДАННЫХ

С целью выявления влияния априорных данных на качество фильтрации, а также анализа чувствительности их к способу речеобразования, были проведены различные эксперименты. Анализ результатов фильтрации зашумленного РС заданного диктора показал, что при отношении сигнал-шум 0 дб. алгоритмы фильтрации могут быть неработоспособными при пренебрежении априорными данными о его голосе. Алгоритм совместной фильтрации РС и оценки сообщения обеспечивает увеличение отношения сигнал-шум, но не очень значительно. Лучшие результаты получаются при совместном оценивании параметров, фильтрации и обнаружении РС. Такая обработка обеспечивает выигрыш в отношении сигнал-шум на 6-9 дб. при вероятности ошибки обнаружения 10^{-2}.

Коротко изложим результаты анализа чувствительности априорных данных к способу речеобразования, смысловому содержанию и индивидуальным особенностям голосов различных дикторов. Некоторые аспекты этой задачи рассмотрены в [5]. Эксперимент состоял в следующем. Подбирались фонограммы (обучающие выборки) разных дикторов. Для каждой фонограммы оценивались траектории сообщений $\hat{a}_{k,t}$, $k=\overline{1,N_\varphi}$, где N_φ – число фонограмм. Для установления факта однородности различных траекторий производилась проверка гипотезы о принадлежности всех

траекторий генеральной совокупности. Если гипотеза справедлива, то различия в траекториях оцениваемых сообщений считаются незначительными и, следовательно, они однородны. В противном случае неоднородны. Это происходит тогда, когда априорные данные чувствительны к способу формирования фонограмм (темп речи, смысловое её содержание, индивидуальность голоса и др.).

Методика проверки истинности гипотезы состоит в следующем. На основе преобразования Фишера по случайным элементам r_{ij} корреляционных матриц $R_a^{(0)}$ и $R_a^{(1)}$ вычисляются новые случайные величины

$$z_{kij}=\frac{1}{2}\cdot\ln[(1+r_{kij})/(1-r_{kij})]. \qquad (7)$$

Определяются усредненные по всем фонограммам величины

$$\bar{z}_{ij}=\sum_{k=1}^{N_\varphi}\alpha_k\cdot z_{kij}\Big/\sum_{k=1}^{N_\varphi}\alpha_k, \qquad (8)$$

где $\alpha_k=M_k/3-3$. Затем формируется случайная величина взвешенного среднеквадратичного отклонения

$$\Delta z_{ij}^2=\sum_{k=1}^{N_\varphi}\alpha_k(z_{kij}-\bar{z}_{ij})^2, \qquad (9)$$

имеющая хи- квадрат распределение с $N_\varphi-1$ степенями свободы. Теперь, если $P\{\chi^2>\Delta z_{ij}^2\}>\alpha$, то принимается гипотеза об однородности различных фонограмм. Здесь α – уровень значимости. Эксперимент проводился на речевом материале удовлетворяющем ГОСТ 16600-72. Параметры обработки РС: частота дискретизации 15 кГц, T_c=15 мс, m=2 и 4.

Анализ экспериментальных данных показал, что априорные данные, полученные для разных дикторов на одной и той же фразе в естественных условиях произнесения с α=0,001 надежно различаются. Способ произнесения и смысловое содержание фонограммы практически не влияют на априорные данные.

В заключении отметим, что проведенный анализ эффективности использования априорных данных подтверждает практическую важность решения поставленной задачи. При этом априорные данные целесообразно формировать для каждого конкретного диктора или группы дикторов со сходными голосами.

ЛИТЕРАТУРА

[1] Ю.Н. Прохоров,"Статистические модели и рекуррентное предсказание речевых сигналов"-Радио и связь: М., 1984.

[2] М.В. Назаров, Ю.Н. Прохоров,"Методы цифровой обработки и передачи речевых сигналов"-Радио и связь: М., 1985.

[3] R. Lee. "Optimal Estimation, Identification and Control." Gamburg, Massachusetts, 1966.

[4] В.Г. Санников, Ю.И. Журавский, Ю.Н.Прохоров,"Формирование банка данных о речи диктора", Тезисы докл. Всесоюзного семинара АРСО-12, Киев-Одесса, 1982.

[5] М.В. Назаров, Ю.Н. Прохоров, Ю.И. Журавский,"Исследование характеристик параметров авторегрессии заданных источников", Тезисы докл. Всесоюзного семинара АРСО-13 Новосибирск, 1984.

SPEAKER INDEPENDENT CLASSIFICATION OF VOWELS AND DIPHTHONGS IN CONTINUOUS SPEECH*

Michael S. Phillips

Computer Science Department
Carnegie Mellon University
Pittsburgh, Pennsylvania 15221
USA

INTRODUCTION

When designing a vowel recognizer for continuous speech, one must consider not only the actual recognition algorithms but also the question of what categories the recognizer should attempt to identify. Should the vowel categories be at a phonemic level or a phonetic level [1] ? For a particular level of labeling, how detailed should the categories be? For example, if a phonetic level of labeling is chosen, which allophones should be grouped together? Another important problem is to find a consistent way of labeling speech at a particular level for training and testing the system.

This paper will describe the current design of the vowel recognizer that is under development and present classification results using two sets of vowel labels. The vowel recognizer is part of the acoustic phonetic recognition module of the CMU DARPA speech understanding system (Adams and Bisiani [?]). The system consists of a signal processing module, an acoustic phonetic recognition module, a word matcher, and a sentence parser. A block diagram of this system can be seen in Figure 1. The acoustic phonetic recognition module is given various representations of the speech signal as input and produces a network representing possible phonetic transcriptions of the speech signal. The network has nodes representing possible segment boundaries and arcs that have lists of labels with associated probabilities.

1. In this paper, the terms "phonemic level" and "phoneme" refer to the speaker's internal representation of the sounds in the lexicon. The term "phonetic level" refers to the actual sound present in the speech signal. For example, speakers may produce the word "children" such that the first vowel would be perceived as an [ah] if listeners were to base their perception only on the acoustic signal (taking into account acoustic context but ignoring expectations from lexical knowledge). This vowel will be considered to be an [ih] at the phonemic level and an [ah] at the phonetic level

2. The term "segment" is used here only to refer to a portion of the speech signal. It is not meant to imply that these are phonetic segments.

3. The term "feature" is being used here as in the pattern recognition literature. It is not intended to mean phonetic feature.

*Supported by grants from DARPA and NSF

THE VOWEL RECOGNIZER

The job of the vowel recognizer is to produce a list of probabilities of vowel labels given begin and end times for a segment [2] of the speech signal. In the system, these begin and end times are produced by the segmentation algorithms. In these classification experiments, the hand transcription boundaries are used as the segment begin and end times. The segment boundaries are not considered to be the boundaries of the relevant information about the vowel since important acoustic information about the identity of the vowel may be present in the vowel's surroundings. These begin and end times are only used to define the portion of speech that the recognizer is to classify.

The vowel recognizer consists of a set of feature [3] measurement algorithms to measure the acoustic properties of the vowel and a multi-dimensional classifier to produce the label probabilities. The set of feature measurement algorithms should capture all of the relevant acoustic information. The feature measurements for the vowels consist mainly of formant measurements at various points in time, formant changes throughout the segment, spectral centers

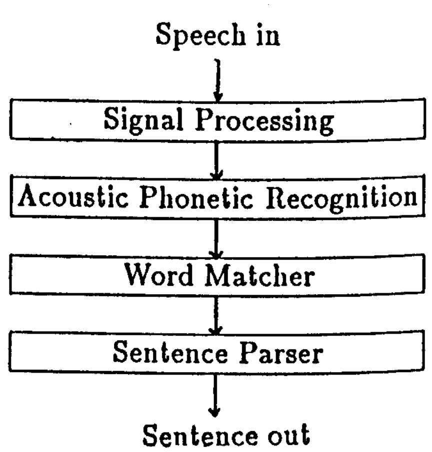


Figure 1: Block diagram of the CMU system.

of gravity measured at various points in time, duration, and average pitch of the segment. Many measurements (for example formant frequencies) are computed in more than one way and it is left to the classifier to decide which is the most reliable way to make a particular measurement for a particular decision. The complete list consists of about 100 feature measurements.

Each segment is represented as a single point in a multi-dimensional space which has a dimension for each of the feature measurements. The job of the classifier is to give the probability of each of the vowel categories given a point in this space (Duda and Hart [2]). The classifier used consists of (n * (n-1))/2 pairwise classifiers where n is the number of vowel categories. The pairwise classifiers are used rather than a single classifier so that the number of dimensions can be as low as possible for each classifier. A single classifier would have to use all of the dimensions that were needed for all labels. Each pairwise classifier is able to use only the features needed for that particular pairwise decision. For example, this allows the system to use a formant tracker specifically designed for front vowels in the [iy] vs [ih] classifier, a formant tracker specifically designed for back vowels in the [aa] vs [ao] classifier, and some spectral centers of gravity for the [iy] vs [aa] classifier.

Each classifier is a two-class, n-dimensional Bayesian classifier where n is the number of feature measurements used for that particular pairwise decision. The classifier assumes a multi-variate Gaussian model of the data samples from each vowel category and given the feature measurements for a segment, assigns probabilities for each vowel category based on that model. Training consists of selecting the best features for each pairwise decision and estimating the parameters of the Gaussian model from a set of training data. Feature selection is done by performing a best-first search through all combinations of available features, using classification performance on a subset of the training data as the criterion for deciding which combination of features is best.

The vowel category probabilities from the pairwise classifiers are combined in the following manner: For each pairwise classifier, the vowel category with the highest probability is given a vote. The probability of the vowel category with the greatest number of votes is the average of the probabilities of that vowel category from each of the pairwise classifiers involving that vowel category. The probability for each other vowel category is the probability of that vowel category from the pairwise classifier for that vowel category versus the vowel category with the greatest number of votes. The probabilities are then normalized so that the sum of the probabilities for all vowel categories equals one.

VOWEL CATEGORIES

In recognition of vowels in continuous speech, the performance of the recognition system will be greatly affected by the choice of vowel categories that the recognizer is attempting to identify. Since the labels given to vowel segments in the training data define the recognizer's models of the vowel categories, the procedure used to obtain these labels is an important factor in determining the system's performance.

The goal of the vowel recognizer is that its decisions should be based on the same information that a human listener would use when making a decision about the vowel. Since the current design of the vowel recognizer works without any top down information from the higher levels of the system, this goal must be altered slightly: the vowel recognizer should use all the information that a human listener uses except any higher level language knowledge.

A listener's perception of a vowel in continuous speech is affected by many factors other than just the acoustic properties of the segment (Rudnickey and Cole [3], Jacob et al. [4]) The neighboring phonemes affect the acoustic realization of the vowel and the vowel affects the realization of the neighboring phonemes. Listeners take into account these local acoustic effects when making a judgement about the identity of the vowel. Listeners are also able to use information from a larger acoustic context (speaking rate, speaker characteristics, etc.) to make a judgement about the vowel. They will also be influenced by their expectation of what the vowel should be given their lexical and semantic knowledge.

The vowel recognition system should take into account as much of the acoustic information as possible to make a decision. This seems to mean that vowel categories on a phonetic level rather than a phonemic level should be used. The acoustic realization of a phoneme depends on higher level rules of the language. Since the vowel recognizer is only able to use acoustic information and not higher level information, it would not be able to map the varying acoustic realizations to the intended phoneme.

It's not clear how to obtain vowel labels at a phonetic level. If listeners are presented with enough of the signal to obtain the complete acoustic context, they will learn what words were spoken and be influenced by the expected phoneme. Alternatively, If short segments of speech are presented, the listeners will only be able to use the local acoustic context. Sentences composed of nonsense words could be used for training and testing the system. This would allow listeners to hear the entire utterance without being having any expectations of the intended vowels. The problem with this approach is that the speakers may have difficulty speaking the sentences in a natural manner.

Since the recognizer must map acoustic information onto probabilities of vowel labels, the labels used for training and testing the system should have as consistent as possible a relationship to the acoustic information. There is ambiguity in phonetic labels (Church [5]). Even if listeners were able to hear vowels in their full acoustic context without being influenced by their phonemic expectation, they would not always agree. Effects such as listeners being influenced by their phonemic expectation or not being presented with the full acoustic context will increase the amount of ambiguity. Since the recognition performance of the system is limited by the amount of ambiguity in the vowel categories in the training and testing data, the labeling procedure should attempt to minimize the amount of additional ambiguity.

LABELING

Two labeling procedures were tried. In one, the people doing the labeling are able to hear the entire sentence and in the other, listeners are given the vowel segment with only a small amount of acoustic context. Both labeling procedures used the same set of labels. This list can be seen in the confusion matrices in Table 3.

The first set of labels used were the hand transcriptions being done for the DARPA speech project's acoustic phonetic database. These transcriptions are made by listening to the utterance, giving a phonetic transcription. running an automatic alignment program, and correcting alignment errors. Besides hearing the whole utterence, transcribersare able to see a spectrogram and other displays and are able to play any section of the utterance. The transcriptions are intended to be phonetic transcriptions but are biased towards the expected vowel in the cases where the realized vowel was ambiguous.

The second set of labels were produced by presenting trained listeners with each vowel segment in its local acoustic context. The segment boundaries were obtained from the hand transcriptions mentioned above. Each vowel segment was first played imbedded in the section of speech starting from the beginning of the transcribed segment before the vowel to the end of the segment after the vowel. After a half second pause, the vowel segment was played in isolation. The listener was able to have these two speech tokens played as many times as necessary. The listener then gave a phonetic label to the vowel with the option of responding with "not sure".

TESTS

The training data for these tests consists of 1000 utterances from 100 (30 female and 70 male) speakers from the DARPA acoustic phonetic database. All of the utterances were labeled both by doing the hand transcriptions and the labeling by listeners described above. The labeling by

	Testing labels Listener labels	Transcription labels
Training labels	(top1/top2/top3)	(top1/top2/top3)
Listener labels	48.3 / 68.7 / 79.3	40.3 / 60.8 / 72.4
Transcription labels	41.4 / 64.3 / 77.1	46.2 / 68.1 / 78.3

Table 1: System performance on the four combinations of training and testing labels. The numbers given are the percent agreement to the testing labels in the top choice, the top two choices, and the top 3 choices of the vowel recognizer.

listeners was done by four listeners (each listener labeled a subset of the 1000 sentences).

The testing data consists of 160 utterances from 20 (6 female and 14 male) speakers. The testing speakers and utterances do not overlap with the training speakers and utterances. The testing data was labeled by the hand transcription and also by three listeners. Each listener labeled all 160 sentences so that listener versus listener agreement could be tested.

The system was trained on both types of labels and tested on both types of labels. The listeners gave a "not sure" label to 3.6% of the segments. These "not sure" labels were ignored during training. For testing the system on listeners' labels, the segments that were given the "not sure" label were automatically relabeled with the label from the hand transcriptions. The results of these tests can be seen in Table 1.

The labels obtained from the three listeners were compared to each other and also to the hand labels. Two comparisons were done: In one, only segments that were not given the "not sure" label by any of the listeners were used. In the other, all segments were used and "not sure" answers were considered to be errors. A summary of These results can be seen in Table 2. Confusion matrices for av-

	Listener 1	Listener 2	Listener 3	Transcriptions
Listener 1	-	64.8/65.8	69.9/65.8	63.0/59.2
Listener 2	64.8/65.8	-	66.9/62.3	59.9/55.1
Listener 3	69.9/65.8	66.9/62.3	-	64.8/62.2

Average Listener versus Listener agreement 67.2/62.7
Average Listener versus transcription agreement 62.5/58.8

Table 2: This table shows the labeling agreement between all combinations of the three listeners and the hand transcriptions. For each entry, the first number is the percentage agreement considering only the segments that no listeners gave a "not sure" label. The second number is the percentage agree for all segment counting "not sure" labels as errors

erage listener versus listener agreement and for average listener versus hand label agreement can be seen in Table 3.

DISCUSSION

From the results in Table 2, it can be seen that the listeners agree with each other 67% of the time and they agree with the hand labels 62% of the time. It seems that there is at least a small difference in these two type of labels. This difference may be due to convention differences between the two types of labels or it may be that the relationship between the acoustic information and the hand labels is less consistent for some distinctions than for the listener's labels. For example, it may be more difficult to make a judgement about the vowel color without being biased by phonemic expectation when listening to the entire utterance. It may also be true that the listening labels are less consistent than the hand labels for some other distinctions. For example, it is likely that it is more difficult to make a decision about vowel reduction in the listening labeling procedure since the listeners were not presented with the entire utterance and do not have access to information about speaking rate and the relative amplitudes of neighboring syllables.

A

	aa	ae	ah	ax	axr	ay	eh	er	ey	ih	ix	iy	aw	ow	oy	ux	uw	uh	ao	ns
aa	337	4	37	6	1	4	1	4	.	.	.	.	1	.	.	.	.	.	39	2
ae	4	216	3	.	.	3	36	.	.	6	.	.	1	.	.	.	.	.	.	3
ah	42	3	191	34	.	4	11	.	.	1	3	.	1	2	.	.	.	6	5	1
ax	12	.	93	140	8	1	6	1	.	7	39	.	.	1	.	.	.	19	4	26
xr	1	.	1	8	62	.	2	67	.	2	2	1	.	.	.	.	.	1	1	5
ay	17	5	8	1	.	165	.	.	11	1	4	1	.	.	.	.	.	.	.	2
eh	4	48	24	16	.	.	325	.	3	28	10	.	.	2	.	.	.	3	2	4
er	5	.	1	.	61	.	.	228	.	1	1	2	.	.	.	.	5	1	2	11
ey	.	4	.	.	.	8	10	.	323	14	2	17	.	.	1	.	.	.	.	7
ih	.	2	17	15	2	.	27	.	12	453	105	55	.	.	.	2	2	8	.	3
ix	.	.	5	57	4	.	11	1	6	151	202	20	.	1	1	2	2	13	.	22
iy	.	.	.	.	.	.	.	1	6	35	7	469	.	.	1	6	.	.	.	4
aw	13	6	2	.	.	.	1	.	.	.	.	.	35	.	.	.	.	.	1	.
ow	3	.	1	3	.	.	.	.	.	.	.	.	2	32	.	.	1	1	14	2
oy	.	.	.	.	.	.	.	.	1	.	.	.	.	.	22	.	.	.	1	.
ux	.	.	.	.	.	.	.	.	.	6	.	7	.	.	.	7	39	.	.	1
uw	.	.	.	.	.	.	.	.	.	4	.	1	.	2	.	42	78	3	.	.
uh	2	.	17	43	2	.	1	1	1	8	10	.	.	.	.	5	15	63	6	.
ao	30	.	3	1	.	.	.	1	.	1	.	.	2	4	1	.	1	2	125	.
ns	15	7	19	33	21	2	21	2	11	30	39	18	3	5	.	3	7	8	7	31
	aa	ae	ah	ax	axr	ay	eh	er	ey	ih	ix	iy	aw	ow	oy	ux	uw	uh	ao	ns

B

	aa	ae	ah	ax	axr	ay	eh	er	ey	ih	ix	iy	aw	ow	oy	ux	uw	uh	ao	ns
aa	231	2	19	3	1	1	.	2	.	.	.	.	2	1	.	.	.	.	10	1
ae	3	226	2	2	.	.	54	.	4	1	.	.	5	.	.	.	.	.	.	4
ah	33	.	194	34	1	1	10	.	.	.	1	.	.	3	.	.	.	13	7	6
ax	15	2	62	163	3	1	14	.	.	16	55	.	.	10	.	.	2	35	14	34
axr	6	.	5	14	98	.	4	95	.	3	10	3	.	.	.	.	.	6	3	29
ay	29	9	7	1	.	168	.	.	3	.	.	.	1	.	.	.	.	.	.	4
eh	2	22	18	3	.	3	264	3	4	24	5	.	.	1	.	.	.	1	2	11
er	6	.	2	5	35	.	5	196	.	6	.	.	.	1	.	3	5	1	3	15
ey	.	5	.	.	.	20	2	1	311	12	2	8	.	.	.	.	.	.	.	5
ih	.	5	9	14	2	.	20	1	10	316	45	19	1	1	.	1	1	10	1	3
ix	.	7	32	110	11	5	82	3	19	276	297	50	.	.	2	2	1	26	3	52
iy	.	2	.	.	.	2	.	3	27	58	30	475	.	.	1	6	.	2	.	15
aw	8	3	1	.	.	.	3	.	.	.	.	.	35	1	.	.	.	.	2	1
ow	4	.	3	3	.	.	.	.	.	2	.	.	3	33	.	.	1	3	18	2
oy	.	.	.	.	.	.	.	.	1	.	.	.	.	.	22	.	.	.	2	2
ux	.	.	.	.	.	.	.	2	1	9	10	5	.	.	.	35	65	6	.	2
uw	.	.	.	.	.	.	1	1	.	2	4	.	.	2	.	16	65	14	.	9
uh	.	.	3	3	1	.	.	.	.	.	2	.	.	.	.	.	.	33	.	3
ao	124	.	6	1	.	.	1	.	.	.	.	.	4	1	.	.	.	1	124	5
ns	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
	aa	ae	ah	ax	axr	ay	eh	er	ey	ih	ix	iy	aw	ow	oy	ux	uw	uh	ao	ns

Table 3: Confusion matrices for average listener versus listener comparison (a) and average listener versus hand transcription (b). In (b) the row is the hand transcription label and the column is the listener label.

When trained and tested on the hand labels, the system's first choice accuracy is 46% and when trained and tested on the listeners' labels, the accuracy is 48%. A larger difference in performance can be seen comparing testing the system on the same type of labels as it was trained on versus testing the system on the other set of labels. Again this could either be explained by a convention difference or by a difference in the types of inconsistencies in the two labeling procedures.

It certainly seems that there is a large amount of ambiguity in the vowel categories being used for the vowel recognizer. The upper limit to the performance of the vowel recognizer is the amount of ambiguity present in the mapping from the acoustic information to the vowel labels. Since the listeners only agree with each other 65% of the time, this is the upper limit for the vowel recognizer performance if it is trained and tested on these labels. Obtaining labels that have a more consistent relationship to the acoustic information either by redefining the vowel categories or by developing a better labeling procedure should directly improve the performance of the vowel recognizer. From the system performance data, it seems that the two labeling procedures investigated so far have approximately equivalent amounts of ambiguity. If some distinctions are made more consistently with one procedure than the other, perhaps a better labeling procedure would combine the best aspects of both.

REFERENCES

[1] Adams, D. and Bisiani, R., The Carnegie-Mellon University Distributed Recognition System. Speech Technology, Mar/Apr 1986, 14-23.

[2] Duda, R. and Hart, P., Pattern Classification and Scene Analysis. John Wiley and Sons. 1973.

[3] Rudnicky, A. and Cole, R., Effect of Subsequent Context on Syllable Perception. Journal of Experimental Psychology, 4(4):638-647. 1978.

[4] Jacob, B. et al., The Effect of Language Familiarity on Vowel Discrimination. Presented at 100th meeting of the Acoustical Society of America, Los Angeles California, November 1980.

[5] Church, K., Phrase-Structure Parsing: A Method for Taking Advantage of Allophonic Constraints. PhD. Thesis, MIT, 1983.

SPEAKER-INDEPENDENT SPEECH-RECOGNITION USING ALLOPHONES

K. Bartkova
Institut de la Communication Parlée
Institut National Polytechnique de Grenoble
46, Avenue Félix Viallet
38000 GRENOBLE, FRANCE.

D. Jouvet
Centre National d'Etudes des Télécommunications
Route de Trégastel
22300 LANNION, FRANCE.

ABSTRACT

This study concerns the determination of the allophones that are necessary for achieving a good recognition of the French numbers by a speech recognition system based on a Markov modelling approach. The allophones have been distinguished, for the vowels, by the formant transitions at the "onset" and at the "offset", and for the consonants, by their phonetical characterization.

For this specific application, using an average of 2 allophones by phoneme and a few "clusters", we achieved 94.9% correct recognition rate on the whole numbers, for 13 speakers that were not in the training set.

INTRODUCTION

A speaker-independent speech-recognition system has to deal with all the possible acoustical realizations of the words in the vocabulary. The variations result from various speakers, different possible pronunciations and coarticulation effects. The recognition system we used [1], based on a Markov modelling approach, can handle part of these variations through the automatic training procedure. However, the basic units, used to describe the words (usually phonemes), have different acoustical realizations depending on the context. If one uses a specific acoustical model for each phoneme in each context, the total number of necessary models would be pretty large. But, for any phoneme, several contexts may have nearly the same influence on its acoustical realization. So a good tradeoff, between accuracy and complexity, is to use different acoustical models, for a given phoneme, only when the context influence is different enough.

That is the reasons why we are studying the allophones as each of them corresponds to a particular acoustical realization. It is worthvile mentioning that this study concerns only a specific application, namely the French numbers between 0 and 999, and thus has no pretention to be a full theoretical research of the French allophones. Nevertheless, the set of allophones determined in this study may be extended as needed to fit a new vocabulary. The French numbers use nearly all the vowels and half of the consonants of the French language. The strict syntax and the limited vocabulary restrict the number of contexts for each phoneme, thus, we were able to conduct a full study of the different contexts for this specific application.

After a description of the data base, this paper details the different realizations of the phonemes. For the vowels, we used mainly the transition of the formants, and for the consonants, their phonetical characterizations. We end the paper by an application of the allophones in a speech recognition system.

DATA BASE

The data base contains about 3000 French numbers between 0 and 999. They were recorded from 34 adult speakers (22 men and 12 women). All the speakers have a "standard" pronunciation, except one having a strong regional accent.

The table lists the different phonemes of the data base. For typographical reasons we denote the phonemes by one or two ascii characters, and we specified here the standard phonetical meaning when different from the notation used.

Vowels	Oral	i, ei(e), ai(ɛ), a, o(ɔ), au(o), ou(u), eu(ø), oe(œ), e(ə).
	Nasal	an(ɑ̃), in(ɛ̃), un(œ̃), on(ɔ̃)
Consonants	Plosive	d, t, k.
	Fricative	v, z, f, s.
	Liquid	r.
	Nasal	n.
	Semivowel	w, y(ɥ).

This study, concerning the determination of the allophones, was conducted using the spectrograms of the data in association with the pitch and the waveform.

VOWELS ALLOPHONES

One of the main acoustical realizations of the context influence on the vowel is the transition of the formants at the "onset" and at the "offset". For practical reasons, related to the implementation of the speech recognition system, we will treat separately the consonantic influence, the pause influence, the possible devoicing and the case of adjacent vowels.

Consonantic Influence

From the locus theory [2, 3], which explains the transition of the formants at the "onset" or the "offset" of the vowels by the point of articulation of the adjacent consonant, we defined 6 classes for the consonants. We grouped together the apico-dental and the predorso-alveolar contexts because the transition of the formants they induce are very similar [3].

Labial	f, v	(labio-dental)
Dental	t, d, n	(apico-dental)
	s, z	(predorso-alveolar)
Velar	k	
Labio-palatal	y	
Labio-velar	w	
Uvular	r	

Instead of measuring by degrees the displacement of the articulation point, or the aperture [4], during the realization of the vowel, we will characterize the allophones by their full context. The next table reports the vowels of the data base and the contexts in which they occur. Each row corresponds to a left context, and each column to a right context. "Lpal" stands for labio-palatal, and "Lvel" for labio-velar. In order to represent all the positions we add the "pause" and "vowel" contexts. They will be treated later on.

Left \ Right	Labi	Dental	Lpal	Vela	Uvul	Paus	Vowe
Labial		in	in	in		in	in
Dental	an	ai,an,eu	an	an	ei	an	an
	oe	i,in,ou		in	o	eu	ei
Lpal.		i					
Lvel.		a				a	
Velar		a,an,in			a		
Uvular	au	ai, au	au	au		au	au
	oe	an					
Pause		on				un	
Vowel		on				un	

Taking each vowel in each possible context would define a full set of allophones for this application. However, one can also define a subset by grouping together for each vowel some contexts which have nearly the same influence. We should point out that this grouping is different from one vowel to another. For example, for the oral vowel /au/ we can put together the right contexts "velar", "labial" and "pause"; but for the nasal vowel /an/ we will have to keep separate the realizations corresponding to the "velar" and "labio-palatal" contexts which induce important transitions on the formants.

Pause Influence

In general 3 different realizations are possible for the "onset" or the "offset" of a vowel when adjacent to a pause. Just after a pause, we can have a glottal stop, a synchronized or an aspirated (devoiced) beginning. Just before a pause, we can have a glottal stop, a synchronized or a devoiced ending. A synchronized beginning or ending corresponds to a progressive rising or falling of the pitch and of the intensity showing a synchronization between the vocal cords vibrations and the velum and the articulators movements. A devoiced beginning or ending results from a partial forward or backward assimilation, the pause having the same effect as a voiceless context.

Vowel devoicing

The voicing feature appears to be rather robust for the vowels. Only one context was strong enough in our data base to devoice a whole vowel. This context was /s.w_s/ for the vowel /a/ in the word "60" (/s.w.a.s.an.t/). After a devoicing of the /w/, the vowel /a/, surrounded by voiceless consonants loses its voicing feature and becomes coarticulated with the surrounding noise.

Adjacent vowels

For 2 adjacent vowels, belonging to different words, we noticed the following realizations: For **unstressed vowels** the transition of the formants is smooth and uninterrupted. For **stressed vowels** either a short pause (50 to 200 ms) appears between the vowels and they may start or end by a glottal stop, or the transition is realized by a glottalized vocalic portion having a low pitch.

Summary

Because of the implementation in the speech recognition system we group together the "pause" and the "vowel" contexts. The pause does not induce formant transitions, and the transitions between adjacent vowels are handled by specific acoustical models. In order to obtain a good representation of the various transitions of the formants we had to define an average of 2 allophones by vowel. The number of allophones used for each vowel, reported in the following table, does not take into account the pause influence and the possible devoicing.

Vowel	i	ei	ai	a	o	au	ou	oe	eu	an	on	in	un
Alloph.	2	2	2	3	1	3	1	2	1	4	1	4	1

SUPRASEGMENTAL INFORMATION

As this speech recognition system does not use pitch information, and also because for such short sentences the pitch is not a useful syntaxic hierarchical cue, the only suprasegmental information we have studied is the segmental duration. The importance of the vocalic duration is justified by the facts that, besides an obvious correlation between word and phoneme durations, the degree of perturbation of the formants by the context is strongly related to the length of the vowel. Also, the knowledge of the minimal duration is useful for designing the acoustical models. For these reasons we have started a statistical analysis of the segment durations.

The vowels before a pause are longer than the same in other positions. This agree with the fact that, in French, the stress is on the last syllable of the sense-group [5,6], which often corresponds for our application to the whole numbers. One of the most important acoustical realizations of the stressed syllables is the longer duration of the vocalic nucleus. In a closed syllable, the influence of the following consonant on the vocalic duration agree with previous studies [7, 8].

The vowel duration in a non final syllable, therefore unstressed position, is strongly correlated with the duration of the sense-group. However the duration of a stressed vowel is independent of this influence. For example, the /a/ appearing in the third syllable of a 6 syllables sense-group lasts 54 ms as regards to 82 ms when being in the unstressed syllable of a 2 syllables word. But, when the /a/ appears in a stressed syllable, it lasts 164 ms in a 6 syllables group even followed by a shortening consonant such as /t/ in French, compared to 144 ms, ceteris paribus, in a 2 syllables word.

NEUTRAL VOWEL - SCHWA

The neutral vowel should be treated like a possible occurence place rather than an acoustical realization pattern. Theoretically, in French, at a slow speaking rate in a careful articulation manner, it is possible to pronounce a schwa at the end of every isolated word ending by a consonant. However, for connected words such as the numbers, this neutral vowel may be pronounced at the end of each of the individual words. For example, whether the schwa (e) is pronounced or not, implies 4 different theoretical patterns for a sequence like "55" (/s.in.k.an.t.(e).s.in.k.(e)/).

For a correct identification of the neutral vowel, one needs to use suprasegmental information such as the vocalic duration. The duration seems to be the more appropriate cue for differentiating the schwa from the vowels /oe/ and /eu/. For example, the duration of the schwa before a pause was always very short compared to the duration of the previous stressed vowel.

CONSONANTS ALLOPHONES

The different realizations of the consonants, are first described using phonetical characteristics such as nasalization, labialization, etc, as modifiers applied to the "standard" realization. After that, we treat the case of the epenthetic sounds and the voicing feature.

Allophonic characterization

Nasalization: This concerns the stop consonants after a nasal vowel. The voiced stop /d/ may, by a forward coarticulation effect, become partly or completely nasalized. For the voiceless stops, a nasal consonant may be realized before it or even replace it.

Palatalization: This concerns the stop consonants in a right labio-palatal context, or before a palatal, anterior vowel.

Labialization: This concerns the fricatives followed by a labio-velar semivowel, or preceeded by a rounded posterior vowel.

Vocalization: This concerns the voiced fricative /v/ and the liquid /r/ in some intervocalic positions (for example /oe.v.in/ in "80" and /a.r.an/ in "40").

Fricatization: The unvoiced realization of /r/ is in a strict sense a fricative [9]. The devoicing, usually due to an adjacent voiceless consonant, may occur, for some speakers, even in an intervocalic context.

Rolled: This concerns, in our data, only the unvoiced /r/ after the voiceless stop consonant /t/. This realization is produced by a flapping (quasi occlusion) between the back of the tongue and the "velo-uvular" region.

Tense: For these data, the consonant duration vary a lot in two positions: first or last consonant of a sense-group adjacent to a pause. Some studies [10] note an increase in the tension of the articulators, the vocal cords and the velum for the initial position of a sense-group and for the stressed syllable of the group. We will denote as "tense" the corresponding realization of the consonants. This characteristic does not correspond to the feature "tense" as defined in some classical theories of segmental phonology [11], but rather defines some consonantic realizations appearing in specific contexts.

An initial voiced consonant may also have a very short duration, and even vanish, in which case the only remaining cues are the formant transitions at the "onset" of the following vowel (for example /y/ in /y.i.t/ or /v/ in /v.in/). For these reasons we have to define, in an initial position, just after a pause, 2 allophones with different acoustical realizations and different durations for the fricative /v/ and the semivowel /y/, one corresponding to a "standard" pronunciation and the other to the "tense" realization. At the end of a sense-group, in a stressed syllable, the VOT of the unvoiced stops, when followed immediatedly by a pause have the same realizations as "tense" consonants for some speakers (important high frequency noise and a longer VOT).

Speaking rate and epenthetic sounds

For some speakers having a rather slow speaking rate we notice the realization of 2 epenthetic sounds, one consonantic and one vocalic. An unvoiced consonantic "closure" is realized in a context where the nasal consonant is preceeded by an unvoiced consonant; this occurs for the consonant /n/ preceeded by the voiceless stop /t/ or the devoiced fricative /z/. A neutral vowel (schwa) may occur when the voiced realization of /r/ is followed by a voiced consonant.

Voicing feature

The voicing feature, for the consonants, is often inaccurate and is strongly influenced by the context. In fact, its modification, due to coarticulations effects, appears to be the same for most of the consonants: stops, fricatives and semi-vowels. For voiced consonants, the pause has the same influence as an unvoiced context, and thus implies a partial or total devoicing by a forward or backward assimilation. The following table gives, for the specified contexts the consonants for which the voicing feature may be modified:

Consonants	Left context	Right context
d, z, v.	Pause	Vowel
z.	Vowel	Pause
y, w.	Unvoiced consonant	Vowel
t, k.	Vowel	Vowel

Summary

The following table lists for each characteristics the consonants that are affected, and the contexts that induce this modification by a forward or backward assimilation. The "*" denotes an irrelevant context (ie anything).

Characteristics	Consonants	Contexts
Nasalization	t, d, k.	/an/,/in/ — *
Palatalization	t.	* — /y/
	d.	* — /i/
Labialization	s.	* — /w/
	z.	/ou/,/on/ — *
Vocalization	r, v.	Vowel — Vowel
"Rolled"	r.	/t/ — *
Fricatization	r.	/t/ — *
	r.	Vowel — Vowel
Tense	v, y.	Pause — *
	t, k (VOT)	* — Pause

RECOGNITION TESTS

We applied this study to the speaker independent recognition of the French numbers between 0 and 999. For this recognition test we used a data base of 26 speakers (14 men and 12 women), each speaker having recorded the 10 digits, 50 random numbers between 00 and 99 and 50 between 000 and 999. Half of this data base was used in the study of the acoustical realizations. This data base, containing about 2900 numbers, was separated into 2 parts. The data from 13 speakers were used for training the model parameters, and the data from the 13 other speakers were used for measuring the recognition performances in a speaker independent mode. The acoustical parameters used are the Mel frequency cepstrum coefficients, plus the total energy and its temporal variation. They are computed every 20 ms (frame rate) using the energy in 24 Mel filters; the bandpass of the signal being 6.4 hKz.

The reference point, for measuring the improvement due to the allophones, is a phonetic based model, in which the words are described as sequences of phonemes, each of them being represented by the same acoustical model, independently of the context. However, because of strong coarticulations effects, the sequences /t.r/, /v.a/ and /y.i/ were considered as basic units and thus were represented by a single acoustical model. Using this description, we achieved 93.1 correct recognition rate on the whole numbers for the testing set. Using an average of 2 allophones by phoneme, introducing specific models to handle transitions between adjacent vowels, and keeping the 3 "clusters" mentionned above, we achieved 94.9% correct recognition rate on the same data base, thus reducing the error rate by 25%.

CONCLUSION

This paper shows that a good description of the vocabulary improves the performances of a speech recognition system. As the coarticulations and the different pronunciations are predicted, the acoustical models have just to take into account the variations due to the various speakers. However, it seems that to correctly predict all the coarticulations, it would be necessary to consider, besides the immediate context, the individualities of the speakers and also the speaking rate of the current sense-group. The set of allophones, defined for this specific application, can easily be extended to fit new vocabularies.

Although the current version of our speech recognition system cannot handle segment duration information, we noticed that the duration is an important cue for differentiating a final devoiced /z/ from /s/. An extra cue for identifying a final devoiced /z/ is the realization of a schwa after the fricative.

BIBLIOGRAPHY

[1] D. Jouvet, J. Monné, D. Dubois: "A new network-based speaker-independent connected-word recognition system"; IEEE proc ICASSP 1986, Tokyo, pp 1109-1112, April 1986.

[2] P. Delattre, A.M. Lieberman, F. C. Cooper: "Acoustics loci and transitional cues for consonants"; JASA, No 27, pp 769-773, 1955.

[3] E. Emerit: "Nouvelle contribution à la théorie des locus - Première partie" Phonetica, Vol 30, No 1, pp 1-31, 1974.

[4] M. Rossi, Y. Nishinuma, G. Mercier: "Indices acoustiques multilocuteurs et indépendance du contexte pour la reconnaissance automatique de la parole"; Speech communication, North-Holland, pp 215-217.

[5] P. Delattre: "Durée vocalique et consonnes subséquentes"; Le Maître Phonétique, 67, 3-ième série, London, 1939.

[6] F. Dell: "L'accentuation dans les phrases en français" in "Les représentations en phonologie" (F. Dell, D. Hirst, J.R. Vergnaud), Paris, Herman, 1984.

[7] A. Di Cristo: "De la microprosodie à l'intonosyntaxe"; Thèse de doctorat d'Etat, Université de Provence, 1978.

[8] K. Bartkova, C. Sorin: "Predictive Model of Segmental Duration in French"; (109-th ASA Meeting), JASA, Suppl 1, Vol 77, p S54, Spring 1985.

[9] F. Lonchamp: "Phonétique et phonologie"; Formation au traitement de la parole, Fascicule 1, Institut National Polytechnique de Grenoble, Grenoble, 1986.

[10] J. Vaissière: "Variance and invariance at the word level" in "Invariance and variability in speech processes", edited by J.S. Perkell and D.H. Klatt, Hillsdale, New-Jersey, London.

[11] R. Jakobson, C. Gunnar, M. Fant, M. Halle: "Preliminaries to speech analysis: The distinctive features and their correlates"; The MIT Press, Cambridge, Massachusetts, 1969.

ТАБЛИЧНЫЙ МЕТОД ВЫДЕЛЕНИЯ ПРИЗНАКОВ РЕЧЕВОГО СИГНАЛА И ПОФОНЕМНОЕ РАСПОЗНАВАНИЕ РЕЧИ

ЛЮДОВИК ЕВГЕНИЙ КУЗЬМИЧ

Институт кибернетики им. В.М.Глушкова АН УССР
Киев, СССР 252207

АННОТАЦИЯ

Поэтапное устранение избыточности из квантованных параметров описания речевого сигнала, а также учет контекста, то есть, окружения каждого интервала анализа направлены в конечном итоге на однозначное указание значения выделяемого признака, а в случае пофонемного распознавания – номера фонемы.

ВВЕДЕНИЕ

Применяемые в распознавании речи типы первичного описания речевого сигнала представляют собой многомерные векторы, между компонентами которых имеются различного рода зависимости. При этом каждая отдельная компонента несет не очень много информации о выделяемом признаке. Информация как бы размазана по компонентам, причем значительная часть информации о значении выделяемого признака содержится в контексте, то есть, в компонентах первичного описания сигнала на соседних интервалах. Естественно в связи с этим попытаться собрать по возможности всю информацию, отжав "воду" из компонент первичного описания, соответствующих интервалам анализа с их окружением.

Предлагаемый метод включает этап обучения, для которого необходимо накопить и разметить обучающую выборку /ОВ/. Формально размеченная ОВ представляет собой последовательность пар $g(n), v(n)$, $1 \leqslant n \leqslant L$, где $g(n)=\{g_i(n)\}_{i=1}^{m}$ – m-мерный вектор первичного описания речевого сигнала на n-м интервале анализа, а $v(n)$ – значение выделяемого признака v на этом интервале, L – объем выборки в интервалах анализа.

Под выделяемым признаком могут пониматься признаки различного характера: например, признак тон/шум, артикуляционные параметры или номер фонемы. Во всех случаях предполагается, что область значений признака разбита на 2^k подобластей, где $k = 1$ для двоичных признаков и $k = 6$ для номера фонемы являются двумя крайними значениями.

В качестве первичного описания могут использоваться самые различные наборы параметров, например, коэффициенты отражения и интенсивность сигнала или же значения энергии в частотных полосах. Могут применяться также и наборы разнородных параметров.

Обучение имеет целью построение преобразования параметров первичного описания данного интервала анализа и окружающих его в значение признака, приписанное этому интервалу.

Построение требуемого преобразования будет производиться поэтапно. При этом от этапа к этапу будет уменьшаться количество битов, необходимых для хранения описания. На заключительном этапе требуемое для этого количество битов должно совпасть с разрядностью k выделяемого признака.

Возможны два варианта постановки задачи построения преобразования. В одном из них процесс построения преобразования направлен на повышение информационной насыщенности каждого бита описания. Во втором варианте целью каждого этапа наряду со сжатием описания является снижение вероятности ошибки распознавания значения признака по отдельным элементам описания очередного уровня. Именно этот вариант и будет рассмотрен ниже.

УМЕНЬШЕНИЕ ИЗБЫТОЧНОСТИ, ОРИЕНТИРОВАННОЕ НА СНИЖЕНИЕ ЧИСЛА ОШИБОК

Предлагаемый метод излагается на примере выделения признака, который может принимать значения от 0 до 15, то есть, требует для задания четырех бит.

Предположим пока, что все компоненты первичного описания проквантованы на 16 уровней и, таким образом, представлены четырехразрядным двоичным кодом.

Образуем составное восьмиразрядное описание из $g_1(n)$ и $g_2(n)$:

$$g^1_{1,2}(n) = g_1(n)\cdot 16 + g_2(n).$$

Исходя из байесовского критерия, построим решающее правило для оптимального распознавания значения признака $v(n)$ по $g^1_{1,2}(n)$. С этой целью для каждого восьмиразрядного двоичного кода k_8 определим то значение k_4 признака v, которое встречается наиболее часто в паре с k_8:

$$k_4 = f^1(k_8) = \operatorname*{argmax}_{k_4} G\{n \mid g^1_{1,2}(n)=k_8, v(n)=k_4\}.$$

Функция G здесь означает количество элементов в множестве-аргументе.

Теперь можно определить первую компоненту описания второго уровня:

$$g^2_1(n) = f^1(g^1_{1,2}(n)).$$

Аналогичная процедура выполняется над третьей и четвертой компонентами описания первого уровня, в результате получаем $g^2_2(n)$ – вторую компоненту описания второго уровня и т.д. Таким образом мы построим преобразование описания первого уровня в описание второго уровня. При этом мы снизим информационный объем описания вдвое, учтем взаимосвязи между парами компонент и, вероятно, повысим надежность распознавания.

Таким же образом переходим к описаниям третьего и других уровней, до тех пор, пока не получим описания из одной компоненты $g^5(n)$ /предположив, что первичное описание состояло из 16 компонент/.

Далее можно учесть окружение n-го интервала анализа. Для этого описанную выше процедуру склеивания следует применить к $g^5(n)$ и $g^5(n-1)$; в результате получим описание следующего уровня $g^6(n)$. Затем объединяем $g^6(n)$ и $g^6(n+1)$ и преобразуем в $g^7(n)$ и т.д. С каждым этапом мы теперь увеличиваем размер фрагмента, участвующего в формировании решения о значении признака $v(n)$ на n-м интервале анализа. Этот процесс может быть остановлен либо при достижении заданной длины фрагмента, либо в том случае, когда число ошибок при распознавании ОВ перестанет уменьшаться.

Если элементы первичного описания имеют большую разрядность, чем значение признака, на первом этапе следует выполнить процедуру склеивания для каждого такого элемента в отдельности.

Если, напротив, разрядность каких-то компонент описания меньше разрядности значений признака, при построении преобразования склеивания следует объединять большее число компонент, что позволит лучше использовать взаимосвязи между ними.

Рассмотрим теперь случай признака большей разрядности, например, пусть этим признаком будет номер фонемы.

Будем исходить из того, что число фонем не превышает 64, таким образом, значения признака $v(n)$ в этом случае являются шестиразрядными двоичными кодами.

Процедура склеивания потребует в этой ситуации построения таблицы из $2^6 x 2^6 = 4096$ шестиразрядных двоичных чисел. В том случае, когда первичное описание состоит из 16 компонент и когда мы хотим для распознавания на n-м интервале использовать информацию о 4-х предыдущих и 4-х последующих интервалах, потребуется объем памяти 48 К байт.

Для уменьшения затрат памяти желательно научиться выполнять процедуру склеивания таким образом, чтобы в результате получались не шестиразрядные коды описаний следующих уровней, а, например, четырехразрядные.

В связи с этим возникает несколько интересных задач.

1. Требуется построить решающее правило, минимизирующее вероятность ошибки распознавания объектов из N классов, если ответами распознавния могут быть только t классов, $t < N$. При этом задается только число классов t, а какие именно это будут классы также необходимо определить.

2. Требуется склеить шестиразрядные коды в 16 групп так, чтобы объединение этого нового четырехбитового описания с некоторым другим фиксированным шестиразрядным позволило бы распознавать номер класса с минимальной вероятностью ошибки.

Решение любой из указанных задач позволяет заменить преобразование-склейку $2^6 x\ 2^6 \to 2^6$ двумя преобразованиями: $2^6 \to 2^4$ и $2^4 x\ 2^6 \to 2^6$, что дает экономию памяти почти в четыре раза. Ясно, однако, что второй вариант в большей степени способствует повышению надежности распознавания.

Еще большую экономию памяти дало бы решение следующей задачи.

3. Допустим, что имеется два шестиразрядных описания $g_1(n)$ и $g_2(n)$. Требуется найти два склеивающих преобразования $f^1: 2^6 \to 2^r$ и $f^2: 2^6 \to 2^s$, $r+s=m$, $6 \leq m$, таких чтобы составное описание $(f^1(g_1(n)), f^2(g_2(n)))$ позволяло бы распознавать номер класса (фонемы) с минимальной вероятностью ошибки.

В этой задаче r и s также являются искомыми параметрами. Задав $m=6$, мы получим экономию памяти не в 4 раза, как в предыдущем случае, а в десятки раз. При этом, однако, надежность распознавания будет меньше, чем в предыдущем случае.

Приняв $m=10$, мы получим вариант в который задача 2 входит в качестве частного случая. Поэтому здесь мы могли бы рассчитывать на повышение надежности распознавания по сравнению с вариантом 2.

Итак, результатом обучения будет совокупность склеивающих отображений, каждое из которых задается таблицей, содержащей информацию о разбиении множества комбинаций значений параметров на подмножества.

Распознавание с помощью этих таблиц тривиально. Нужно последовательно преобразовывать значения компонент описания в соответствии с этими таблицами; окончательный результат преобразования описания на данном интервале анализа и является ответом распознавания. Необходимыми операциями при этом являются операция объединения двух кодов в один код-аргумент и операция извлечения из соответствующей ячейки таблицы кода-значения параметра описания следующего уровня.

ЗАКЛЮЧЕНИЕ

Предложенный табличный метод способствует накапливанию (концентрации) информации, существенной для распознавания. Процедура распознавания тривиальна и использует лишь две очень простые операции. Учет контекста дает основания надеяться на эффективность табличного метода распознавания. Использование многодикторной ОВ создает предпосылки для распознавания речи многих дикторов без подстройки.

AUTOMATIC RECOGNITION OF WORDS

DIFFERING IN DISTINCTIVE QUANTITY

G. Kuhn and K. Ojamaa

96 Leigh Avenue
Princeton, NJ 08540
USA

Library of Congress
Washington, DC 20540
USA

ABSTRACT

We report the results of an experiment on talker-dependent, connected recognition of 10 Estonian CVCV words that differ in distinctive quantity. The words were spoken, and recognized, in sentence pairs of the form "Did you say (word 1, word 2, word 3)? No, I said (word 4, word 5, word 6)." The test sentences were spoken either at the same rate as the training sentences, or at a much faster rate. Each word was modelled with spectral estimates for four variable-duration states.

The best recognition results obtained on the test words spoken at the training (faster) rate, were 88% (64%) without probabilities or likelihoods of durations or duration ratios, 87% (68%) with likelihoods of durations, and 85% (77%) with likelihoods of duration ratios.

We conclude that speech rate can be a major problem for automatic recognition of these words, and that in these experiments the problem was not completely overcome using ratios of successive state durations.

INTRODUCTION

In the field of automatic speech recognition, there is new interest in implicit [1] and explicit [2,3] modelling of speech state durations. However, unless there is a correction for speech rate, expected state durations may be inappropriate. In languages which use distinctive quantity, like Estonian or Finnish, inappropriate state durations could lead to misrecognition of a large number of words.

In this paper, we report the results of an experiment on automatic recognition of 10 Estonian CVCV words that differ in distinctive quantity. Estonian is described as having three consonant quantities and three vowel quantities: short, long and overlong [4,5,6,7]. Within our vocabulary of 10 Estonian words to be recognized, 4 words participated in 2 two-way quantity contrasts: tee:de-teete and kyde-kuu:de; and 6 words participated in 2 three-way contrasts: toode-toote-too:te and kade-kate-katte.

CORPUS

Speech was recorded while one of the authors (KO) read a prepared text. The text consisted of a randomization of 36 occurrences of each of the 10 words, embedded in 60 repetitions of the sentence pair "Kas sa ütlesid (Did you say) 'word 1, word 2, word 3'? Ei ma ütlesin (No I said) 'word 4, word 5, word 6'". The randomization was constrained so that each word occurred 6 times in each position in each sentence of the pair.

The text was recorded 3 times. In the first two recordings, one sentence pair was spoken every 6 seconds. In the third recording, one sentence pair was spoken every 4 seconds. The first recording was used to train the word models, while the second and third recordings were used for the recognition tests.

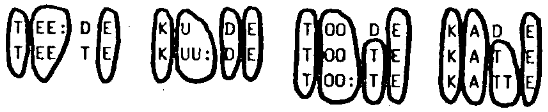
Each recording was digitized at 10000 samples/s. The digitized recordings were parameterized in centisecond frames using a 10-channel, filter-bank spectrum analyzer.

WORD MODELS

We used 15 "word" models, one each for Kas sa, ütlesid, Ei ma, ütlesin, (pause), and the 10 CVCV words. The models for ütlesid and ütlesin had six states. The models for all other words had four states. Each state had an initial segment of fixed duration, a center segment of variable (possibly 0) duration, and a final segment, again of fixed duration. The minimum duration of a state was thus the sum of the durations of its initial and final segments. The minimum durations of the four states in the 10 CVCV words were 3+2, 3+3, 3+2, and 2+3 cs.

The word models were trained using two passes through the training productions. Pass 1 started with DP alignments [8] to the "miniav". The miniav for each word is that training production which has minimum average distance to all training productions of the word. Pass 1 alignments minimized the distance between each training production and the miniav. Means and a covariance matrix were computed over the spectra aligned to each segment of each hand-marked state of the miniav. Pass 2 alignments maximized the probability of the training productions given the Pass 1 means and covariances. Duration estimates (minimum, average, maximum) for each state were produced from the Pass 2 alignments.

In some experimental conditions, spectral estimates were tied across word models, i.e., the weighted average of the means and the weighted average of the outer-product matrices were computed over corresponding segments of the states looped together below:



(Here we refer to the states by name. We feel that this is justified because of the good correspondance over alignments to the miniav). The weights were the number of spectra aligned to each segment. When spectral estimates were tied, there was no spectral difference between the models in word pairs kude-kuu:de, toote-too:te, and kate-katte.

RECOGNITION

The routines for connected recognition computed a spectral match score for the best path through an entire recording [9,10]. That score was the maximum product of the likelihoods of the observed spectra, over all segments of all states of all words on the path. The likelihood of a single spectrum O_t under the continuous multivariate-gaussian probability density function (pdf) for spectral shape in segment j of state i of word w, was

$$L(O_t|j,i,w) = \frac{P(O_t|j,i,w)}{\Sigma_j \Sigma_i \Sigma_w P(O_t|j,i,w)} .$$

The recognition routines used the notion of a "contrast group". Let G(w) be the contrast group for word w, i.e., the group of words including word w that we expected to be confusable under a pure spectral match score. Kas sa, Ei ma and (pause) were each assigned to a one-word group. Ütlesid and ütlesin were assigned to a two-word group. The 10 CVCV words were assigned to four contrast groups, one for each V_1: /e/, /u/, /o/ or /a/.

The recognition options were:

1) expanded range of state durations;
2) restricted word order;
3) independent probabilities of state durations;
4) independent likelihoods of state durations given the contrast group;
5) multivariate likelihood of state durations given the contrast group;
6) independent likelihoods of a pair of state duration ratios given the contrast group;
7) independent likelihoods of a second pair of state duration ratios given the contrast group;
8) multivariate likelihood of the second pair of state duration ratios given the contrast group.

With expanded state durations, durations in the range $0.5*min_{i,w}$ through $1.5*max_{i,w}$ were permitted.

With restricted word order, Kas sa could only follow (pause); ütlesid could only follow Kas sa; Ei ma could only follow (pause), ütlesin could only follow Ei ma; while the other 10 words and (pause) could follow one another any number of times.

With independent probabilities of state durations, the spectral score for each possible duration of each state was multiplied by $P(d_i|i,w)$. $P(d_i|i,w)$ is the probability of duration d_i in state i of word w, under a discrete binomial state duration pdf parameterized by $(min_{i,w}, average_{i,w}, max_{i,w})$.

With independent likelihoods of state durations given the contrast group, the spectral score for each possible duration of each state was multiplied by

$$L(d_i|i,w,G(w)) = \frac{P(d_i|i,w)}{\Sigma_{m \in G(w)} P(d_i|i,m)} .$$

With the multivariate likelihood of state durations given the contrast group, the spectral score for each word w was multiplied by the tri-variate gaussian $L(d_{S-2},d_{S-1},d_S|w,G(W))$, where S is the number of states in word w.

With independent likelihoods of a pair of state duration ratios, the spectral score for each word w was multiplied by $\Pi_r L(ratio_r|w,G(w))$, r=1,2. The underlying duration ratio pdf's, $P(ratio_r|w)$, were discrete binomials parameterized by the (min,expected,max) values of $ratio_r$. The first pair of duration ratios tested [11] was

$$ratio_1 = d_{S-2}/(d_{S-2} + d_{S-1})$$
$$ratio_2 = d_{S-2}/(d_{S-2} + d_S) .$$

The second pair of duration ratios tested [12] was

$$ratio_1 = d_{S-2}/(d_{S-2} + d_{S-1})$$
$$ratio_3 = (d_{S-2}+d_{S-1})/(d_{S-2}+d_{S-1}+d_S) .$$

With the multivariate likelihood of the second pair of state duration ratios given the contrast group, the spectral score for each word was multiplied by the bi-variate gaussian $L(ratio_1,ratio_3|w,G(w))$.

RESULTS

Boxes are drawn on the confusion matrix in Table 1. Let the count in the boxes divided by the count in the 10 rows be a "similarity score" (these words were at least recognized as a word in the same contrast group). Then this confusion matrix shows how a recognition score of 88% and a similarity score of 99.4% was obtained when a baseline system was run on the 6s/pair test recording. The baseline system used the observed range of state durations, separate spectral models, unrestricted word order, and a path score based only on the spectral match.

Figure 1 gives recognition results in terms of recognition scores on each test recording, and average similarity score over the two test recordings. The curve of recognition scores for the 6s/pair test recording is labelled "6". The curve of recognition scores for the 4s/pair test recording is labelled "4". The curve of average similarity scores is labelled "SIM".

Under conditions 0-3 in Figure 1, the baseline system was used (condition 0), or the baseline system modified by three cumulative changes: expanded range of durations (condition 1), tied models (condition 2), and restricted word order (condition 3).

Not surprisingly, both the recognition score for the 4s/pair recording, and the average similarity score, improved with the expanded range of durations.

The recognition score for both recordings decreased with the tied models, because there was no difference between the models in word pairs kude-kuu:de, toote-too:te, and kate-katte, so the routines always chose the first listed word of each pair. However, the average similarity score increased with the tied models, from 97.2% to 98.2%.

Restricted word order did not significantly affect the recognition or similarity scores.

Conditions 4-9 of Figure 1 used expanded durations, tied models, and restricted word order. Conditions 4-6 used recognition options 3-5, respectively. Conditions 7-9 used recognition options 6-8, respectively.

DISCUSSION AND CONCLUSION

As Figure 1 shows, the best recognition results obtained on the test words spoken at the training (faster) rate, were 88% (64%) without probabilities or likelihoods of durations or duration ratios, 87% (68%) with likelihoods of durations, and 85% (77%) with likelihoods of duration ratios.

Figure 2 is a plot of $L(ratio_1|w,G(w))$ for the CVCV contrast groups (from top to bottom) with V_1 = /e/, /u/, /o/ or /a/. Figure 3 is the analogous plot for $ratio_3$. The solid curves are for the models made from the training productions. The dashed curves are for models made post hoc from the 4s/pair productions. As modelled, the $ratio_1$ contrast between toote and too:te was neutralized at the faster rate of speech.

Figure 4 is a scatter plot of the values of $ratio_1$ and $ratio_3$ observed while modelling the CVCV words in the training recording. Figure 5 is the analogous plot for the 4s/pair test recording. Polar coordinates were used for these plots, i.e., the radius is $ratio_3$, and the angle is $ratio_1*\pi/2$. Assuming independence, quantity contrast boundaries lie along radii or along rays.

Figure 6 is a scatter plot of the values of the durations of V_1 and C_2 observed while modelling the CVCV words of the training recording. Figure 7 is the analogous plot for the 4s/pair test recording. The minimum permitted state durations were apparently somewhat long for the 4s/pair recording.

We conclude that speech rate can be a major problem for automatic recognition of these words, and that in these experiments the problem was not completely overcome using ratios of successive state durations.

REFERENCES

[1] R.K. Moore, M.J. Russell and M.J. Tomlinson, "Locally constrained dynamic programming in automatic speech recognition", Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing, 1982, pp. 1270-1273.
[2] T.H. Crystal and A.S. House, "Characterization and modelling of speech-segment durations", Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing, 1986, pp. 2791-2794.
[3] M.J. Russell and A.E. Cook, "Experimental evaluation of duration modelling techniques for automatic speech recognition", to appear in Proc. IEEE Intl. Conf. Acoustics, Speech and Signal Processing, 1987.
[4] P. Ariste, "A quantitative language", Proc. Third Intl. Cong. Phonetic Sciences, 1938, pp. 276-280.
[5] G. Liiv, "On the quantity and quality of Estonian vowels of three phonological degrees of length", Proc. Fourth Intl. Cong. Phonetic Sciences, 1962, pp. 682-687.
[6] I. Lehiste, "Temporal Compensation in a quantity language", Ohio State University Working Papers in Linguistics, 12, 1972, pp. 53-67.
[7] A. Eek, "Estonian quantity: notes on the perception of duration", Estonian Papers in Phonetics, 1979, pp. 5-29.
[8] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimisation for spoken word recognition", IEEE Trans. ASSP-26, February 1978, pp. 43-49.
[9] T.K. Vintsyuk, "Element-wise recognition of continuous speech consisting of words of a given vocabulary", Kibernetika, 7, 1971, pp. 133-143.
[10] J.S. Bridle, M.D. Brown and R.M. Chamberlain, "Continuous connected word reconition using whole word templates", Radio & Electronic Engineer, 53, 1983, pp. 167-173.
[11] U. Lippus, "Prosody analysis and speech recognition strategies: some implications concerning Estonian", Estonian Papers in Phonetics, 1978, pp. 56-62.
[12] K. Ojamaa, "Temporal aspects of phonological quantity in Estonian", Ph.D. Thesis, Univ. of Connecticut, 1976.

	0	1	2	3	4	5	6	7	8	9
0 TEE:DE	36									
1 TEETE		36								
2 KUDE			36							
3 KUU:DE				36						
4 TOODE.					36					
5 TOOTE					2	17	17			
6 TOO:TE						11	25			
7 KADE						1		35		
8 KATE						1		1	25	9
9 KATTE										36

Table 1. Confusion matrix obtained when a baseline system was run on the 6s/pair test recording.

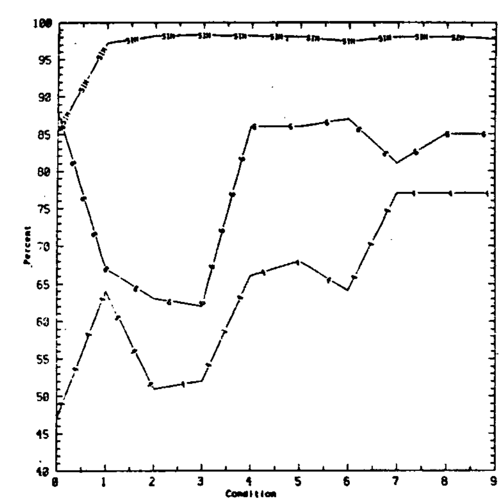


Figure 1. Recognition and similarity scores as a function of experimental condition.

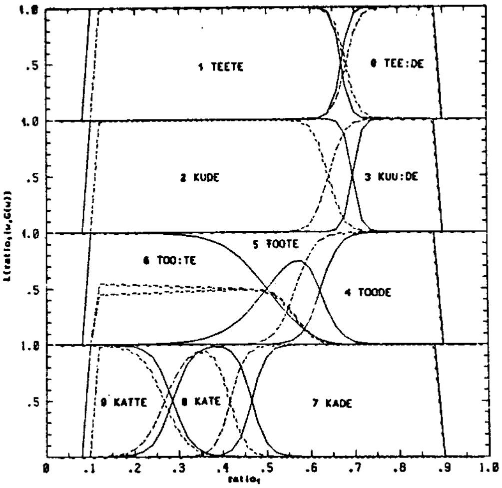


Figure 2. $L(ratio_1|w,G(w))$ for $w=6,8$ and training (solid line) or 4s/pair (dashed) models.

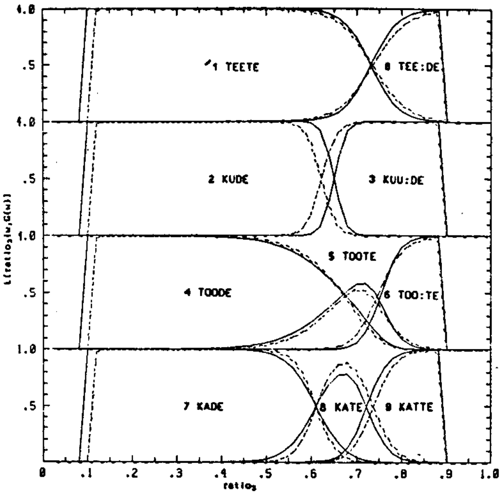


Figure 3. $L(ratio_3|w,G(w))$ for $w=6,8$ and training (solid line) or 4s/pair (dashed) models.

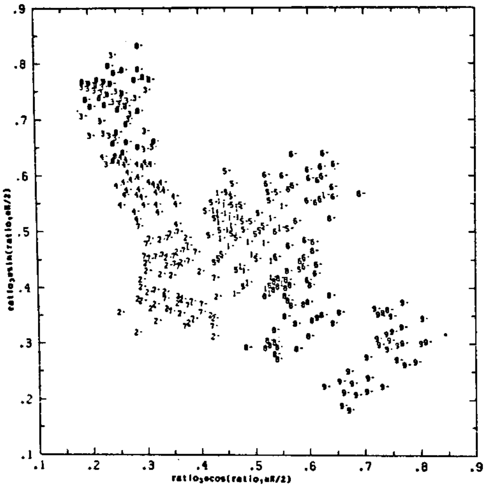


Figure 4. Values of $ratio_1$ and $ratio_3$ observed while modelling the CVCV words in the training recording. Radius=$ratio_3$. Angle=$ratio_1 * \pi/2$.

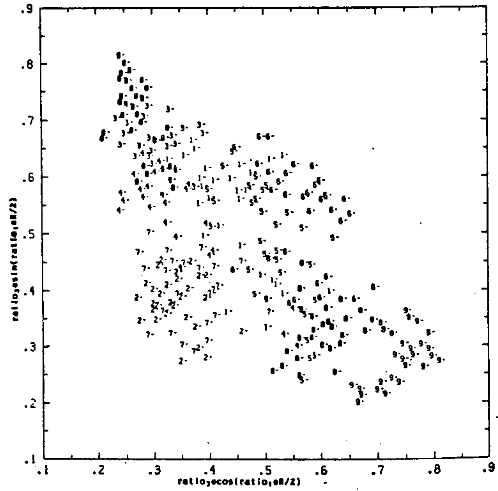


Figure 5. Values of $ratio_1$ and $ratio_3$ observed while modelling the CVCV words in the 4s/pair recording. Radius=$ratio_3$. Angle=$ratio_1 * \pi/2$.

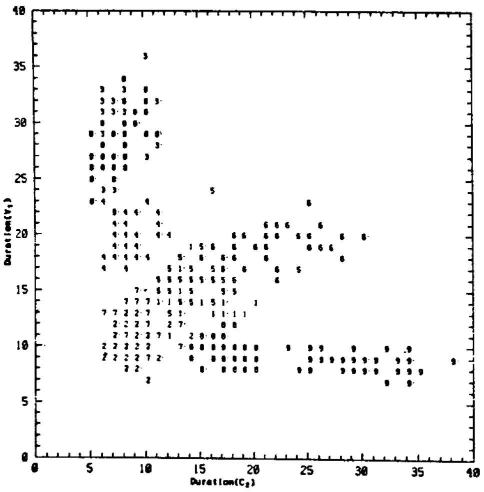


Figure 6. Durations of V_1 and C_2 observed while modelling the productions of the CVCV words in the training recording.

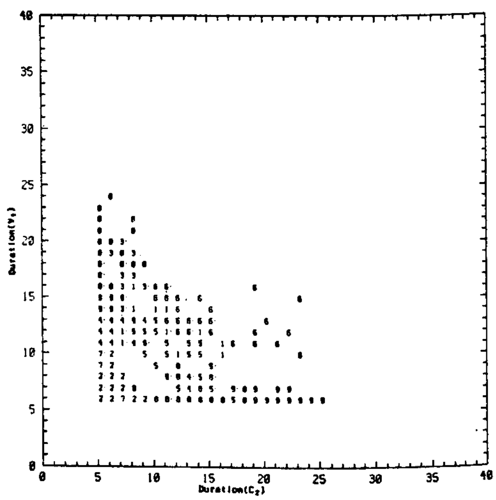


Figure 7. Durations of V_1 and C_2 observed while modelling the productions of the CVCV words in the 4s/pair recording.

COMPUTER RECOGNITION OF ISOLATED WORDS IN FIXED LENGTH FEATURE SPACE

ALGIMANTAS RUDŽIONIS

Speech Research Laboratory
Kaunas Polytechnical Institute
Kaunas, Lithuania, USSR 233028

ABSTRACT

After the bounaaries of separate words were defined, each word was divided linearly into K_S static and K_D dynamic segments. With a fixed number of spectral components L, arbitrary reference words have equal volumes. Several vocabularies of digits and of 100 geographic names were read, and recognition accuracy was estimated in relation to K_S, K_D, L and to the number of repetitions R. It is shown, that with a satisfactory training, even small computers can recognize about 100 words with a 1 to 2% error rate without any devices of increasing their operation speed.

I. INTRODUCTION

A good number of studies on the main parameters of speech recognition systems, and to the main factors determining their success have appeared recently /1-4/.We find such studies as /3, 4/ revealing the difficulties encountered in attemps to evaluate the advantages of different parameter descriptions of signals, as well as of different methods of comparison.
Some of the recent works consider comparison of words divided into fixed number of segments /5,6/. This approach ensures considerably higher recognition speeds,as compared to the non-linear time mode.Perhaps because of the insufficient local similarities of words, the advantages of dynamic transformations of time function are not always exploited.
The present study was aimed at an evaluation of speaker dependent recognition. when separate words are described by fixed numbers of static and dynamic segments of the speech signal, and when the reference of any word is of the same dimension.This mode of description opens new potential ways of word comparison, including polydimensional rating. We consider here the most simple, but not the less important parameters: number of static K_S and dynamic K_D segments, number of spectral components L and effect of training as number of repetitions R.
Our results on several vocabularies of Russian words suggest a continous decrease of the error rate with increasing K_S, K_D, L, R. In particular, which $K_S \sim 8$, $K_D \sim 4$, $L \sim 8$ and satisfactory training, on-line recognition of random vocabularies of about 100 words is possible without any increase in the processing speed, at the error rate of 1 to 2 %.

2. GENERAL REMARKS

The suggested approach has the undoubtful advantage of speed. The effects of small values of K_S, K_D, L, R on its reliability is to be estimated yet. Intuition sugthat higher values of these parameters should extract more information from the speech signal. But we are also interested in the least admissible values and expect significant effects of the vocabulary content. Let us now consider some special aspects of the problem.

2.1. <u>Extration of Features from Speech Signals</u>

Normal logarithmic spectral from the filter bank were chosen as primary features. A word was first described by sequence of spectral vectors $S_K(\ell)$ every T seconds, where $K=1,2,\dots,K_W$ number of spectral vector; K_W – duration of a word expressed by the number of vectors, and $\ell=1,2,\dots,L$ number of a spectral component. For a word divided into K_S static segments, its division points φ_s are

$$\varphi(s) = INT[1+(s-1)(K_w-1)/K_s] \qquad (1)$$

INT stands for a whole part, $S=1,2,\dots K_S+1$. Static segments $F_S(\ell)$ are formed by averaging spectral vector in intervals $K=\varphi_s,\dots,\varphi_{s+1}$

$$F_s(\ell) = \frac{1}{\varphi_{s+1}-\varphi_s+1}\sum_{k=\varphi_s}^{\varphi_{s+1}} S_k(\ell) \qquad (2)$$

The filter analyser used contained L=24 filters, discributed on a semi-logarithmic scale. Dimensional variation of features (L) over the frequency scale was performed either by averaging subsequency spectral frequency components, or by choosing a certain part of component say, in the range of the telephone channel, which is further denoted by L_T.

2.2. <u>Introduction of Dynamic Segments</u>

A certain part of useful information in recognizing speech signals may be gained from changes of the signal spectrum. Its most simple estimation is through the spectrum dynamics

$$\tilde{S}_k(\ell) = \sum_{j=1}^{J} |S_{k+j}(\ell) - S_k(\ell)| \qquad (3)$$

where $K=1,\dots,K_W-J$. For the K_D dynamic segments $F_D(\ell)$, in (2) $S_K(\ell)$ is replaced by $\tilde{S}_K(\ell)$. One of the other possible estimations of spectrum dynamics is weighting by linear functions. But its result was not significantly better, so that (3) was used thanks to its simplicity. The absolute value of (3) has an effect of limitation, but we used it successfully to denote preservation of feature values inside one byte, without any additional operation. Vectors $S_K(\ell)$ were followed every 10 ms with assumed J=3. Size of a reference word was in our case equal to $L(K_S\ K_D)$ bytes and did not depend on the duration K_W of a word. Recognition was carried out by Euclidian distance minimum.

2.3. <u>The Level of Training</u>

The often applied one-time reading of a vocabulary as a means of training for isolated words recognition systems suffers a high degree of randomization of reference words. As a main parameter in our study we chose the estimate of the training set, that is the number of repetitions R for a reference word. The level of training is significantly dependent on the distance measure and on the dimension of the reference word. From practical considerations we consider here only small values of R. On the other hand, a proper level of training can reveal the information efficiency of separate features, and the application limits of the suggested approach.

2.4. <u>Vocabularies Studied</u>

The resultant recognition accuracy is significantly influenced by the vocabulary length and especially by its content. To combine practical interest with fundamental solutions, we selected the following vocabularies: w_1 – names of 100 towns in the Soviet Union, included as a demonstration of the recognition accuracy on vocabularies of medium length; w_2 – Russian digits from zero to thousand (38 words); w_3 – Russian digits from zero to nine (10 words); w_4 – Russian digits from eleven to twenty (10 words). Vocabularies w_3 and w_4 constitute parts of vocabulary w_2. Vocabulary w_3 consists of most frequent in computer recognition words, and vocabulary w_4 is a most difficult vocabulary, because all its words have a common stressed second part.

3. EXPERIMENTAL RESULTS

Most of the results refer to a single speaker. Hardware consisted of a dynamic microphone, a spectral analyser and a microcomputer. Each word was repeated several times, but direct recognition was estimated repeatedly on the second reading, until a statistically significant relation was found. Where possible, several parallel reference vocabularies were formulated. The number of control outputs for each test point was from 500 to 8000.

The main studied parameters were: K_S, K_D – numbers of static and dynamic segments; L – number of spectral components (in the telephone channel range L_T); R – number of training repetitions per word; N – vocabulary length; B – number of bits for a vector component of a reference or of an input; N_C – number of control inputs; T – period of following the spectral vectors from filter analyser ms.

<u>Number of static segments.</u> Fig.1 shows recognition accuracy for different K_S. A significant effect of the vocabulary structure is evident, Fig.1b, combined with the effect of the level of training Fig.1a. Vocabulary w_4 is by far more difficult, than vocabulary w_3, because in w_4 information is concentrated in the first unstressed parts. Vocabulary w_2, which is nearly four times longer, is nearly as difficult as vocabulary w_4 for R=I, but its recognition accuracy improves largely with the number of repetitions, Fig.Ia. In general, recognition accuracy is unsatisfactory at K_S 8.

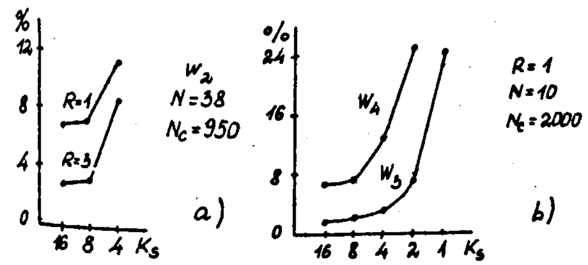


Fig.1. Recognition errors as a function of the number of static segments. (L_T=8, T=10)

<u>Dimension of spectral representation</u>. Typical results in Fig.2 support the expected increase of the recognition accuracy at larger L.

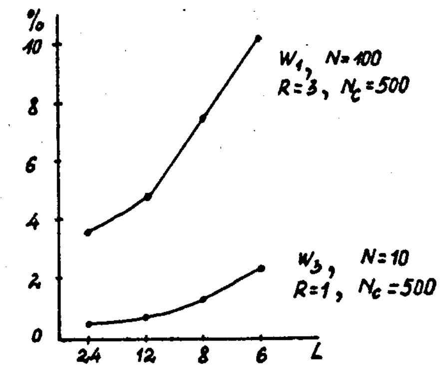


Fig.2. Recognition errors under the effect of spectral resolution (K_S=4, T=20).

<u>Level of system training</u>. The effect of R as also estimated on vocabularies w_3 and w_4 which described by K_S=4 static segments. The error rate in vocabulary w_4 was never lower than $p\approx 7\%$ because of its incomplete information, even at levels of training as high as R=I00. Vocabulary w_3 was nicely recognized at R=3 (Fig.3.

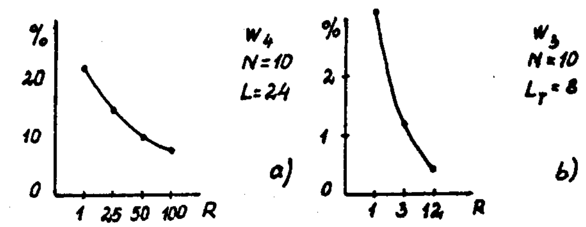


Fig.3. Recognition errors of w_3 and w_4 at different levels of training (K_S=4, T=20).

<u>Effect of spectrum dynamics</u>.

Introduction of dynamic segments was shown in 2.2. Let the words be presented K_S=6 static and K_D=4 dynamic segments. A proper level of training was ensured by numerous repetitions (R=20). It follows from table I, that dynamic segments carry 2 to 4 times less information than static ones, but combined application of both gives 3 times error rates, than static segments only. Note also, that with a satisfactory training, the accuracy of recognition is nearly independent of the vocabulary length.

Table 1. Recognition errors of three vocabularies from their static dynamic and mixed segments, % (L_T=8, R=20, $N_C \approx$ 1000)

K_S	K_D	w_1 N=100	w_2 N=38	w_4 N=10
6	-	6,5	5,3	5,6
-	4	15,2	19,5	11,6
6	4	1,8	1,6	2,2

<u>Presentation accuracy of reference words.</u> The volume of a reference word depends on the number of bits, which are given for each component of the vectors sequences. On the other hand one might expect that the more reliable reference words are also

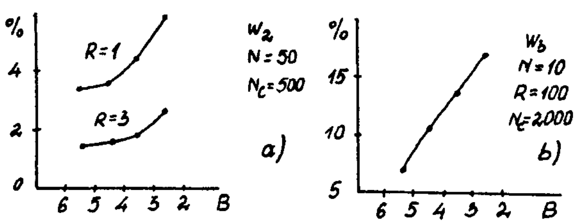


Fig.4. Recognition error rate of the bit number in the reference sequence (L=24, K_S=4).

less sensitive to the accuracy of their presentation. Fig.4 shows the recognition errors of 50 words (vocabulary w_2 plus 12 control words) and of vocabulary w_4 of ten words, when the reference words and the input words were represented by a different number of bits (B). The lower curve, Fig. 4a represents reference words of increased reliability thanks to higher R. Here a decrease of B down to 3 bit has no significant influence on the result. Vocabulary w_4 is highly sensitive to a decrease B of even at R=100.

4.CONCLUSIONS

1. Recognition of words in finite spaces of features with small reference sequencies (8 to 16 segments) and proper levels of training can be suggested for vocabularies of up to 100 words and medium processing speed.
2. Introduction of dynamic segments considerable (3-fold) descrease of the erros rate. A continous decrease of the error rate was observed with increasing K_S, K_D, L and R.

/1/ N.R.Dixon, H.F.Silverman, "What are the significant variables in dynamic programming for discrete utterance recognition?", ICASSP 81, pp.728-731.
/2/ A.Weibel and Yegnanarayana, "Comparative study of Nonlinear Time Warping Techniques in Isolated Word Speech Recognition Systems", IEEE Trans.ASSP-31, pp.1582-1586, No.6, Dez.,1983.
/3/ B.A.Dautrich, L.R.Rabiner, T.B.Martin, "On the Use of Filter Bank Features for Isolated Word Recognition",ICASSP 83, pp.1061-1064.
/4/ N.Noncerino, F.Soong, L.Rabiner, D. Klatt,"Comparative Study of Several Distortion Measures for Speech Recognition", ICASSP 85, pp.25-28.
/5/ D.Burton, "Applying Matrix Quantization to Isolated Word Recognition", ICASSP 85, pp.29-32.
/6/ H.Iizuka, "Speaker Independent Telephone Speech Recognition", ICASSP 85, pp.842-845.

THE PRINCIPLES OF PHONETICAL STRUCTURING OF VOCABULARY FOR SPEECH RECOGNITION SYSTEM

Valeria Kuznetsova

Department of Philology, Moscow State University
Moscow, USSR, 119899

ABSTRACT

In the present paper the problem of decoding the results of the first stage of speech recognition into vocabulary units is discussed. The open syllable is proposed as the basic element for such decoding. The final decision is made with consideration both lexic and phonetic context. The context function is carried out by specially organized vocabulary module in the system.

INTRODUCTION

Lately the problem of mapping the results of preliminary acoustic analysis onto linguistic units draws great attention of different researchers. This problem is very important both for the description of the model of human speech perception and for developing the system of automatic speech recognition and understanding.

The purpose of the present paper is to suggest the solution of this problem in relation to the system of automatic speech recognition. Not discussing in detail the problem of human speech perception, we adopt the following starting-point hypothesis:

I. The decision about signals phonetic content is made for elements corresponding to syllables.
2. Until the content is correlated with the semantic meaning of the unit it is considered to be preliminary and is represented by a limited set of variants or by generalized phonetic content.
3. To arrive at the final interpretation of the signal (to correlate it with some vocabulary unit and to define its phonetic composition more accurately) multifold strategy is implemented on the basis of the information supplied by phonetic and higher levels of analysis.

The present paper deals with the problem of the phonetic structuring of vocabulary module, so without taking into consideration higher levels of linguistic analysis we'll describe some possible model of transition from signal representation (in terms of the first stage alphabet) to vocabulary units.

PHONETIC SYLLABLE RECOGNITION

The basic element of our recognition model is an open syllable. The selection of this unit is supported both by the acoustic-phonetic literature data regarding it to be the minimal unit of speech perception and production /I/-/5/ and by the possibility of automatic segmentation of the results of the first stage recognition into elements corresponding to open syllables /5/, /6/.

The results of the first stage recognition presented in /6/ were used to test the model's reliability. Signal, corresponding to syllable, automatically having been singled out and recorded in terms of the first stage alphabet (FSA) is compared with syllable sample (SS) from the system's memory. Each SS is correlated to phonetic syllable. Thus the result of the first stage recognition goes into the input of the given submodule while in the output there are syllables in phonetic transcription.

These SS were designed on the basis of analysis of the results of the first stage recognition of the definite system with regard to possible within-syllable coarticulation and the duration of syllable's constituents. Thus SS are in their nature idealized, generalized concept of the results of the first stage recognition and are recorded in terms of FSA. The SS set is determined by the requirements put on the recognition system vocabulary. It is rather small in case of limited vocabularies. For evolving systems of automatic speech recognition with extensive and unlimited vocabularies the SS set must be compiled with regard to syllable statistics. The existing syllable statistics for Russian speech /7/, /8/ do not fully answer the requirements of this problem as they are received on the basis of idealized transcription of written texts. Contrary to the statement in /I/ syllables constituting these statistics cover no more than 60% of different type oral texts, as it was shown in our experiment. Thus taking the statistics presented in /8/, as the starting point we are now compiling a fuller statistics that would comprise up to I000 open syllables revealed from the recordings of different types of oral texts. This statistics would supply the basis for SS set of the speech recognition system with extensive vocabulary in which every syllable would get SS representation. The model was tested with SS set of I00 syllables and vocabulary of 200 words.

The syllable corresponding to signal is selected by means of comparing the entering signal to each SS and is determined by minimal istance between them. The distance is measured y means of consecutive comparison of each si-nal constituent to elements of the sample with he help of the matrix of phonetic distance MPD) stored in the system's memory. MPD comp-ises conventional distances between elements f FSA constituting the signal and SS.

We used the following technique for crea-ting MPD. Each element of FSA corresponds to a certain set of acoustic features. It can be characterized by presence/absence of some feat-ure and the strength of its manifestation (e.g. absence of fundamental frequency is characteri-stic of unvoiced consonants; by different de-grees of its manifestation along with several other features, voiced obstruents, sonorants and vowels are distinguished).

The difference between FSA elements regar-ding each feature was estimated by assigning certain marks to them. The results of our ana-lysis /9/, /I0/ of reliability of these featu-res in recognition were taking into considera-tion. The distance between reliable features was given a higher mark, while the distance be-tween less reliable ones was given a low mark, that is the scales of distances were not line-ar. Thus the scales were made not for elements of alphabet, but for the features by which these elements are characterized. The summarized dis-tance between the constituents of the compared features of FSA elements was put into MPD. In the process in a number of cases frequent subs-titution of elements of the alphabet in the si-gnal or complete absence of such substitution was taken into consideration.

The technique described above can be prese-nted in the following way: if M and N are ele-ments of FSA, and M is characterized by the set of features $/x_1, \dots, x_i/$ while $N - /x_1', \dots, x_i'/$ then

$$R_{M,N} = r_{(x_1,x_1')} + \dots + r_{(x_i,x_i')} + k_{M,N}$$

where $R_{M,N}$ is the distance included in MPD, $r\ /x_i, x_i'/$ - the distance in the scale for each feature, $k_{M,N}$ - correction coefficient of sub-stitution frequency of elements in the definite recognition system.

We have distinguished and scaled the fol-lowing acoustic features:

I. Fundamental frequency
2. Presence of formant structure and degree of its manifestation
3. Intensity
4. Main area of energy concentration
5. FII frequency
6. FI frequency

These acoustic features are highly analo-gous to syllable contrasts described by L.Bon-darko /I/. The main difference here is the abs-ence of durational contrast in our scales. It is impossible to introduce this feature into MPD because the decision is made about each ti-me segment of the signal, and not about segment corresponding to some phonetic unit (whether sound or syllable). The coefficient of compari-son between signal duration and sample duration is introduced into algorithm for calculation of distance between the result of the first stage recognition and SS. It seems interesting to compare our data with those obtained on the basis of /I/. By means of the technique descri-bed above we have constructed MPDI on the basis of the scales corresponding to syllable cont-rasts description in /I/. Naturally the absolu-te distance rarely coincide as singled out fea-tures do not match completely, although some general tendency in the sequence of elements of the alphabet which are arranged according to the degree of closeness to each element can be observed. We are planning to compare the effi-ciency of the matrix in the recognition system.

The results of comparing the signal with SS set allow us to put forward a preliminary hypothesis about some syllable corresponding to the certain signal. As it was mentioned above such decision is represented by either a set of syllables with minimal distances from the sig-nal (in our case 3 minimal distances were taken into consideration) or by symbolic recording of the syllable, reflecting generalized phonetic content (e.g. TA - a syllable consisting of un-voiced stop and non-front vowel). Whether a set of variants or a generalized content would be selected for syllable recording depends on the signal's character (the degree of manifestation of features that allow us to define some conc-rete sound with greater or lesser precision) or on its distance from the sample. Such attitude seems quite reasonable as not always in the si-gnal there are acoustic cues that would allow us to correlate it with some definite sound, syllable or even word /II/, /I2/.

THE STRUCTURING AND USE OF THE VOCABULARY

As the result of the program for comparing the signals of the first stage recognition with SS each word is represented in form of open-sy-llables' string. This fact determines the cha-racter of phonetic description of the vocabula-ry. The constructing of the vocabulary can be divided into 2 stages.

On the first stage lexical units in the form close to idealized phonetic transcription are recorded as strings of open syllables. On the second stage pronounciational variants and the most frequent substitutions in recognition being singled out in the preliminary analysis are included into transcribed word recording. If limited vocabulary is used on this stage it is advisable to set apart possible quasihomonyms such as [KAPAT'-KATAT'] (in this example the vowels of initial syllable are identical while the consonants at the beginning of the second syllable are phonetically very similar and pra-ctically undistinguishable in the process of recognition or are distinguished irregularly).

Each transcribed recording is correlated with corresponding word or words and in the ca-se of reliable syllable recognition we get spe-lling of the words on the vocabulary output.

The program for syllable joining compares all possible strings of syllable-candidates with those recorded in the vocabulary and cor-responding to real words. These equivalents are then recorded into spelling and sent to the output. Variants of input strings of syllables that do not correspond to any vocabulary unit are eliminated. This program imitates the role of the lexical context in phonetic recognition. In some cases it's possible that a whole group of syllable strings would correspond to vocabu-lary units, thus we'll get 2 or more words at the output. During program approbation such ca-ses were rather few and the number of words at the output didn't exceed 3. This can be expla-ined by the small size of the vocabulary. Theo-retically the number of variants for the se-lected number of syllable candidates /3/ is 3^X, where X stands for number of syllables in the given word. We suppose that in such cases the elimination of extra variants is possible on a higher level of analysis and it corresponds to the role of syntactic, pragmatic and semantic context in speech perception.

A more complex case is presented by the situation when some syllables are identified incorrectly and none of the strings of syllab-les at the input of the vocabulary module cor-responds to the vocabulary units. In this case multiple strategy of word search must be imple-mented. This strategy must be based on some fa-ctors that determine identification of the sig-nal with a lexical unit and its segment compo-sition /in other words the strategy is phone-tic-context dependent/. The number of syllables in a word, stress position, rhythmical structu-re of a word as a whole, basic /most reliable in the process of recognition/ syllables, ini-tial syllables, consonant clusters can be named as such phonetic factors here. The type of the selected factors and their number cause the vo-cabulary structure, the determining of absolu-tely reliable factors cause in its turn the strategy of word search in general: consequent search beginning with subvocabularies, composed according to absolutely reliable word characte-ristics /in respect of the process of recogni-tion/ and onto subvocabularies based on less reliable phonetic word characteristics, were only candidates selected with the help of "re-liable" subvocabularies are taken into conside-ration. As the first stage recognition results don't allow us to consider every selected fac-tor absolutely reliable, one has to turn to pa-rallel word search strategy, although it's a rather complicated procedure.

The vocabulary has the following structure. The vocabulary is recorded in the form of its variants /subvocabularies/, which are organized in accordance with the selected factors. Some subvocabulary is derived into parts comprising similar rhythmic structures, another is orien-ted at the basic syllables and so on. Strings of syllable candidates which have no correspon-ding lexical elements in the main vocabulary are entered into all these subvocabularies and there are selected every subvocabulary word ca-ndidates identical /in the structural characte-ristic of a certain subvocabulary/ to the ente-red string. Word candidates are entered into the analyser which in the output delivers words present in all registers. If no such words can be identified word candidates with the highest marks are selected. For this purpose to each of the subvocabularies a certain rank is assigned according to the reliability of the factor ref-lected in it.

At present we are conducting an experiment aimed at selecting factors used in word recog-nition and defining the degree of their reli-ability. For this purpose the results of syl-lable recognition with false decision were gi-ven to a group of experts, who using the words' phonetic features and unlimited vocabulary put forward some hypothesis about lexical correla-tion of these results. The group consists of 4 linguists who can theoretically ground their decisions. The data thus obtained are of preli-minary character but it should be pointed out that the experts pay attention to the words' rhythmical structure and to the segment compo-sition of stressed and initial syllables.

VOCABULARY MODULE

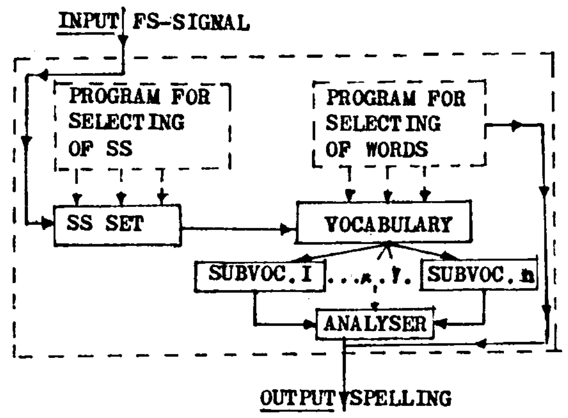


Fig. I
The generalized scheme of the vocabulary module work

With the use of limited vocabulary and pragmatically oriented recognition system the strategy of word prediction can be used /I2/. In this case the reliability of recognition of syllables and the probability of their substi-tution must be taken into account. The vocabu-lary must be built in the form of a matrix ref-lecting the consecutive member of each syllable in a word and the search is conducted through the vocabulary beginning with the most reliable syllable to its possible left or right neigh-bours. If the supposed neighbouring syllables coincide with the result of the recognition or if they are not contradictory to it /that is, are included into the register of possible sub-stitutions or have the generalized phonetic content/ the search is conducted further on. On the basis of some given text pragmatics seman-tically and syntactically oriented subvocabula-ries can be selected to conduct the search whi-le the sequence of entering into each subvoca-bulary will be determined by the previously re-cognized words.

This model has no computer-program realization, yet we present it here in accordance with the concept that in speech recognition as well as in human speech perception it's impossible to limit oneself to one particular strategy. The final decision can be made on the basis of identification of the word image as a whole, on the basis of the analysis of the factors /phonetical as well as relating to other levels of analysis/ determining this image, on the basis of prediction of syllables and larger units /words and word combinations/ by limiting the communicative vocabulary according to the pragmatic content of the text.

MODEL'S EXPERIMENTAL APPROBATION

Model's partial approbation /syllable recognition and word selection in the vocabulary with the help of the vocabulary of basic syllables/ was conducted on the vocabulary of 200 words and a set of I00 syllable samples. Syllable candidates are obtained as the result of the realization of the program for comparing of the first stage signals with SS. Strings formed of syllable candidates are recoded into vocabulary units. As the experiment demonstrates, for a small vocabulary it's sufficient to introduce I or 2 subvocabularies. 3 operational variants of the program are possible: simple joining of recognized syllables, word prediction by means of the subvocabulary of basic syllables, refusing to make final decision in case of false recognition or absence of the basic syllable in the string. For the purpose of limiting the number of analysed strings the syllables undoubtedly falsely recognized /initial and final syllables that got into middle position, middle syllables that got into initial or final position/ are eliminated. The result appears in the spelling form with an index showing the ratio between the number of correctly recognized syllables and the total number of syllables in the word. Below some examples of different variants of the decision are given:

TRANSMITTED	RECOGNIZED SYLLABLES I	2	3	THE RESULTS
maja	/maI	ja/	-	maja
	/ma	mna/	-	2/2
	(maI)	(mə)	-	
vzaimno	/vza	iI	(rvə)	vzaimno
	/za	(/i)	(nə)	2/3
	(za)	(/p'i)	(va)	zaika
				2/3
v des'at'	/vd'eI	s'ə	t'/	v des'at'
	(d'eI)	z'iI	f/	3/3
	/d'eI	(/či)	s'/	des'at'
				3/3

/ma - initial syllable
d'eI - stressed syllable
f/ - final syllable
iI - basic syllable
(maI) - syllable eliminated by the program

The experiment was conducted on computer SM-4 with positive results.

REFERENCES

/I/ Бондарко Л.В. Фонетическое описание языка и фонологическое описание речи. Л., I98I.

/2/ Бондарко Л.В. Слог: правила, интуиция, механизмы. - В кн.: Функциональная просодия текста.М.,1982.

/3/ Бондарко Л.В. Акустические характеристики речи. - В кн.: Слух и речь в норме и патологии /вып. I/. Л., I974.

/4/ Уровни языка в речевой деятельности: к проблеме лингвистического обеспечения автоматического распознавания речи./Под ред. Л.В.Бондарко. - Л., I986.

/5/ Белявский В.М., Светозарова Н.Д. Слоговая фонетика и три фонетики Л.В.Щербы. - В кн.: Теория языка. Методы его исследования и преподавания. Л., I98I.

/6/ Белявский В.М. Автоматическая сегментация слитной речи. - В сб.: IX Всесоюзная акустическая конференция. Тезисы докладов. М.,1977.

/7/ Елкина В.Н., Юдина Л.С. Статистика открытых слогов русской речи. - В сб.: Вычислительные системы, I4. Новосибирск, I964.

/8/ Златоустова Л.В. и др. Алгоритмы преобразования русских орфографических текстов в фонетическую запись. М., I970.

/9/ Москаленко Т.А. Акустический анализ согласных звуков в целях автоматического распознавания русской речи. - В сб.: Автоматическое распознавание слуховых образов: Тезисы докл. и сообщ. АРСО-I4. Каунас, I986.

/I0/ Кузнецова В.Б., Смирнова О.Н. Анализ надежности автоматического распознавания фонетических признаков. - Там же.

/II/ Проблемы и методы экспериментально-фонетического анализа речи. Л., I98I.

/I2/ Кузнецова В.Б. О возможном способе формирования словарных эталонов. - В сб.: Автоматическое распознавание слуховых образов: Тезисы I2-го Всесоюзного семинара АРСО-I2. Киев, I982.

SEARCH FOR OPTIMAL TEACHING PROCEDURE AND WARPING ALGORITHMS FOR ON ISOLATED WORD RECOGNITION DEVICE

K. Vicsi, G. Lugosi, T. Linder

Acoustical Research Laboratory
Hungarian Academy of Sciences

An automatic word recognizer was developed using a preprocessor constructed on the base of a psychoacoustic model. In every 10 ms 5 parameters characterize one word. Different time warping algorithms were examined and made a comparison between them to find the optimum for the recognition accuracy and for the recognition time. Some teaching method was also examined to choose the best for recognition using the simplest teaching procedure.

GRUNDBEGRIFFE EINER THEORIE DER AUSSPRACHENORM

M. W. RAJEWSKI

Philologische Fakultät
Staatliche Lomonossov-Universität zu Moskau
UdSSR, Moskau, 119899

Im untenstehenden Beitrag wird auf die Notwendigkeit der Entwicklung einer Theorie der Aussprachenorm hingewiesen. Da eine Theorie nur mit Hilfe einiger Grundbegriffe entwickelt werden kann, werden folgende Begriffe zu diesem Zwecke vorgeschlagen und erörtert: Aussprachenorm, nichtkodifizierte Aussprachenorm, kodifizierte Aussprachenorm, Bestandteile einer kodifizierten Aussprachenorm, Merkmale einer kodifizierten Aussprachenorm, normative Aussprache und ihre Formen.

Das aktive Interesse für die Sprachnorm, das in den letzten Jahrzehnten in der UdSSR und im Ausland an den Tag gelegt wurde, äußerte sich in zahlreichen Veröffentlichungen allgemeinen Charakters, die sich mit solchen Fragen der Sprachnormtheorie beschäftigten wie Verhältnis von Sprachnorm und Sprachsystem, Stabilität und Variabilität der Sprachnorm u.dgl. Große Aufmerksamkeit wird der Erforschung der einzelsprachlichen Normen in ihrer Entwicklung und ihrem heutigen Zustand geschenkt. Neben der Erforschung verschiedener Aspekte einzelsprachlicher Sprachnormen entwickelt sich auch eine allgemeine Theorie der Sprachnorm /1/. Doch liegt, wie man sich leicht davon vergewissern kann das Hauptaugenmerk auf den grammatikalischen und lexikalischen Normen, während die Aussprachenorm (im folgenden - AN) im Hintergrund bleibt und in rein theoretischer Hinsicht nur beiläufig behandelt wird. Der Grund dafür wird wohl darin zu suchen sein, daß die Bemühungen von Fachleuten auf dem Gebiet einzelsprachlicher Aussprachenormen bis in die jüngste Zeit hinein der Beschreibung und Kodifizierung einer AN sowie der Ermittlung ihrer Variabilitätsgrenzen galten, denn die Erfordernisse des praktischen Sprachunterrichts machten eben diese Aufgaben aktuell. Doch sollten einzelsprachliche Aussprachenormen nicht nur in einschlägigen Nachschlagewerken beschrieben und kodifiziert werden, sondern auch eine theoretische Begründung und Einschätzung im Rahmen der allgemeinen Sprachnormtheorie erfahren. Eine notwendige Voraussetzung dafür bildet die Aufstellung von dazu erforderlichen Grundbegriffen, die auch entsprechend definiert werden müssen. Im folgenden wird der Versuch unternommen, einige Begriffe aufzuzählen und zu erörtern, die m.E. zu den Grundbegriffen einer Theorie der AN gehören sollten.

Die AN bildet einen Teil der Sprachnorm im großen und ganzen. Darum ist es zweckmäßig, die Definition der AN aus der Definition der Sprachnorm schlechthin abzuleiten. Von der von Prof. Dr. O.S. Achmanova vorgeschlagenen Definition der Sprachnorm ausgehend (die Sprachnorm sei "der anerkannte Gebrauch von Sprachmitteln in der Rede, die Gesamtheit von Regeln (Reglementationen), die den Gebrauch von Sprachmitteln in der individuellen Rede ordnen" /2/), kann man die AN als den anerkannten Gebrauch von lautlichen Ausdrucksmitteln einer Sprache oder als die anerkannte Gestaltungsweise der lautlichen Seite einer individuellen Rede definieren. Die Grundlage einer AN bilden natürlich entstandene oder bewußt formulierte Regeln, die den Gebrauch von lautlichen Ausdrucksmitteln einer Sprache in der individuellen sprecherischen Tätigkeit steuern.

Je nach den Funktionen, die verschiedenen Sprachen eigen sind, sind zwei Arten von Aussprachenormen zu unterscheiden: nichtkodifizierte und kodifizierte Aussprachenormen. Eine nichtkodifizierte AN ist die Norm, die keine bewußte Festlegung erfahren hat und nur in mündlicher Überlieferung existiert als eine Gesamtheit von vage als Norm empfundenen Vorschriften und Verboten, die die individuelle Artikulationstätigkeit von Muttersprachlern regeln. Eine nichtkodifizierte AN ist für Orts- und Territorialmundarten, schriftlose Sprachen, sowie Literatursprachen in früheren Perioden ihrer Entwicklung charakteristisch.

Eine kodifizierte AN ist die Norm, die bewußt festgelegt, ausführlich beschrieben und als eine Gesamtheit von Regeln in entsprechenden Nachschlagewerken und Anleitungen dargestellt ist. Also existiert sie für die Muttersprachler als ein von ihnen freiwillig und bewußt akzeptiertes Muster, das als Regulator ihrer artikulatorischen Tätigkeit funktioniert. Die kodifizierte AN bestimmt die Gestaltungs-

der lautlichen Seite der individuellen sprecherischen Tätigkeit auf der Grundlage von bewußt durch Fachleute formulierten und vom Sprecherkollektiv bewußt akzeptierten Regeln, die für den Gebrauch lautlicher Mittel der jeweiligen Sprache maßgeblich sind. Sie ist nur für entwickelte, allseitig normierte Literatursprachen charakteristisch. Dank der Vielfalt ihrer Funktionen sollten solche Sprachen maximal reglementiert sein sowohl in ihrer schriftlichen als auch in ihrer mündlichen Existenzform, was die Aneignung dieser Sprachen und ihre Anwendung als Verständigungsmittel erleichtert. Die Kodifizierung einer AN setzt eine ziemlich hohe Entwicklungsstufe der artikulatorischen Phonetik voraus und stellt die letzte, abschließende Etappe in der Festlegung von Normen einer entwickelten Literatursprache dar.

Die AN einer entwickelten Literatursprache beruht auf einer detaillierten Beschreibung der lautlichen Seite der mündlichen literatursprachlichen Kommunikation von Muttersprachlern. Diese Beschreibung ermöglicht es vor allem, den Lautbestand der jeweiligen Literatursprache in ihrer mündlichen Form, d.h. ihren normativen Lautbestand zu ermitteln. Die Laute der jeweiligen Literatursprache sollen genügend einförmig artikuliert werden, was durch die Befolgung von Regeln ihrer normativen Artikulation gesichert wird.

Doch werden die Laute im Sprechstrom in Gruppen artikuliert, welche Folgen von miteinander nahtlos verbundenen, ja einander durchdringenden akustischen Ereignissen darstellen, die durch die artikulatorische Tätigkeit des Sprechers hervorgebracht werden (eine solche minimale Folge bildet die Silbe). Die artikulatorische Tätigkeit des Sprechers verläuft innerhalb eines zwischen zwei Pausen liegenden Sprechstromabschnittes ununterbrochen. Darum sollen die Artikulationskomplexe, die für das Aussprechen von unmittelbar benachbarten Lauten erforderlich sind, unbedingt nicht nur einander angepaßt werden, sondern auch einander durchdringen, was die notwendige artikulatorische Fusion und die Kontinuität der Lautung ermöglicht. Darum werden die Regeln der normativen Artikulation von isolierten Lauten durch die Regeln ihrer normativen Artikulation im Sprechstrom ergänzt.

Die Befolgung der beiden Regeltypen erlaubt es, 1) eine mehr oder weniger einförmige Aussprache dieser Laute zu erreichen und 2) die Deutlichkeitsschwelle nicht zu überschreiten, unterhalb welcher die Aussprache undeutlich wird.

Endlich enthält die AN einer Literatursprache auch den in orthoepischen Wörterbüchern festgelegten normativen Lautbestand einzelner Wörter. Die Festlegung dieses "idealen phonetischen Bestandes von Wörtern" /3/ vereinheitlicht die Aussprache und Lautung von Wörtern und Wortformen im Sprechstrom. Von Zeit zu Zeit finden im normativen Lautbestand einzelner Wörter Veränderungen statt, die durch die Umverteilung von Lauten im literatursprachlichen Wortschatz hervorgerufen werden. Diese Veränderungen sollen in den nachfolgenden Auflagen von Aussprachewörterbüchern und sonstigen Nachschlagewerken ihre Widerspiegelung finden.

Die AN einer entwickelten Literatursprache, als Gesamtheit ihrer obengenannten Teile betrachtet, weist folgende Merkmale auf:

1) sie ist kodifiziert, d.h. als Gesamtheit von Regeln in einschlägigen Anleitungen festgelegt;

2) sie ist bearbeitet, d.h. sie unterliegt bestimmten Regeln, die das Resultat einer zielgerichteten Tätigkeit von Experten sind, welche die Auswahl der optimalsten Artikulationen und kommunikativ am meisten berechtigten Realisationen einzelner Laute und des normativen Lautbestandes einzelner Wörter bezweckt;

3) sie ist zugänglich, d.h. sie enthält nur solche Regeln, die es allen Muttersprachlern erlauben, sich die AN beim Erlernen der Literatursprache anzueignen;

4) sie ist stabil und veränderlich zugleich, d.h. sie behält ihre Qualität im ganzen im Laufe einer gewissen Zeit unter allmählichen Veränderungen einiger Regeln in der sprecherischen Praxis, was die periodischen Präzisierungen und manchmal sogar eine teilweise Neufestlegung der Norm notwendig macht.

5) sie hat einen sozial bewußten Charakter, d.h. sie ist von den Trägern der jeweiligen Literatursprache als die einzig richtige Gestaltungsweise der lautlichen Seite der literatursprachlichen mündlichen Kommunikation, als Muster, an welches sie ihre artikulatorische Tätigkeit angleichen, bewußt akzeptiert;

6) infolge der bewußten Anerkennung durch die Träger einer Literatursprache ist sie für jeden Muttersprachler in kommunikativen Situationen verbindlich, die die Anwendung der jeweiligen Literatursprache erfordern.

Solche Merkmale der kodifizierten AN wie ihre Kodifiziertheit, ihr Bearbeitetsein, ihre Zugänglichkeit, Stabilität und Veränderlichkeit sind ihr als Regulator der artikulatorischen Tätigkeit beim literatursprachlichen Sprechen immanent eigen. Solche Merkmale aber wie bewußten Charakter und Verbindlichkeit erhält die kodifizierte AN sozusagen von außen im Ergebnis ihrer Billigung und Anerkennung durch die Träger der jeweiligen Literatursprache.

Die Aussprache, die der von den Trägern einer Literatursprache akzeptierten AN entspricht, wird als normative Aussprache (Literaturaussprache, Musteraussprache usw.) bezeichnet. Die normativen Aussprachen von Literatursprachen der Gegenwart existieren mindestens in zwei

Formen. Eine davon ist ihre am meisten verbreitete Standardform, die für alle Sprecher der jeweiligen Literatursprache in kommunikativen Situationen verbindlich ist, in welchen sie sich mit der Notwendigkeit konfrontiert sehen, die Literatursprache in ihrer mündlichen Form anzuwenden. Eben diese Form wird gemeint, wenn man von der Literaturaussprache im eigentlichen Sinne des Wortes oder der Standardaussprache (dem orthoepischen Standard) spricht. Die andere Form der normativen Aussprache ist ihre Höchstform, die eine tadellose Lautung der literatursprachlichen mündlichen Rede sichert. Sie ist vor allem für das bühnengerechte Sprechen der Schauspieler charakteristisch und wird aus diesem Grund als Bühnenaussprache bezeichnet/4/. Die Bühnenaussprache stellt gleichsam die berufsbedingte Form der Literaturaussprache dar.

Die beiden Formen differieren voneinander vor allem durch die Unterschiede im Grad der Konsequenz, mit welchem die Sprecher den normativen Lautbestand von Wörtern einer Literatursprache realisieren, doch können sie voneinander auch in Besonderheiten der Artikulation einzelner Laute abweichen. Dabei ist für die Bühnenaussprache eine strenge Einhaltung der kodifizierten Lautung von Wörtern und ihren Formen charakteristisch, denn nur die maximal genaue Wiedergabe ihres normativen Lautbestandes ermöglicht es den Zuschauern, wie weit entfernt von der Bühne sie auch sitzen mögen, Wörter und Wortgruppen mühelos zu identifizieren und somit alles zu verstehen, was auf der Bühne gesprochen wird. Die Literaturaussprache im eigentlichen Sinne des Wortes (Standardaussprache) läßt eine bestimmte Variabilität des normativen Lautbestandes eizelner Wörter und Wortformen zu, die durch verschiedene Ursachen hervorgerufen werden kann (Sprechtempo, Charakter der kommunikativen Situation, regionale Besonderheiten der Artikulation einzelner Laute, Unterschiede im Entwicklungsgrad orthoepischer Fertigkeiten u. dgl.).

Die normative Aussprache der Literatursprachen der Gegenwart weist in ihrer Standardform oft regionale Abarten auf, was vor allem für Sprachen charakteristisch ist, die auf größeren Territorien gesprochen werden /5/. Die Standardaussprache der polynationalen Literatursprachen (z.B. Englisch, Spanisch, Deutsch) existiert in einer Reihe von nationalen Varianten. Regionale Typen bzw. nationale Varianten der normativen Aussprache einer Literatursprache unterscheiden sich voneinander dadurch, daß jeder bzw. jede von ihnen einige Besonderheiten enthält, die durch die Einwirkung territorialer Dialekte bedingt sind und von Generation zu Generation mündlich überliefert werden.

Von der kodifizierten Standardaussprache heben sich ihre regionalen Abarten ab durch Abweichungen von den normativen Regeln der Artikulation einzelner Laute bzw. durch Verletzungen des normativen Lautbestandes einzelner Wörter. Regionale Abarten der Literaturaussprache im eigentlichen Sinne des Wortes, die in den Grenzen eines Nationalstaates nebeneinander bestehen und verwendet werden, kann man als regionale Typen dieses orthoepischen Standards betrachten.

Nationale Varianten der Standardaussprache polynationaler Literatursprachen, die sich voneinander auch durch Besonderheiten der Artikulation einzelner Laute und Abweichungen im normativen Lautbestand einzelner Wörter unterscheiden, stellen nationale Aussprachenormen polynationaler Literatursprachen dar.

Regionale Typen der Standardaussprache erfahren keine bewußte Festlegung in einschlägigen Anleitungen und orthoepischen Wörterbüchern. Ihre Besonderheiten werden in diesen Nachschlagewerken lediglich erwähnt, wenn sie deren Autoren wichtig erscheinen. Ebenso werden die Wandlungen innerhalb regionaler Typen der Standardaussprache höchstens nur mitberücksichtigt. Dafür aber sind die Wandlungen in der kodifizierten normativen Aussprache - welcher Herkunft sie auch sein mögen (Folgen von spontanen Lautveränderungen in bestimmten Positionen oder Veränderungen in der sozialen Zusammensetzung der Träger einer Literatursprache, Konkurrenz von territorialen Aussprachevarianten usw.) - ein Gegenstand ständiger Beobachtung von Fachleuten, da es ihre Aufgabe ist, jede kodifizierte AN von Zeit zu Zeit mit der fortschreitenden Entwicklung der jeweiligen Literatursprache in Einklang zu bringen.

Mit Rücksicht auf alle vorhergehenden Ausführungen könnte folgende präzisierte Definition einer kodifizierten AN vorgeschlagen werden: unter einer kodifizierten AN ist eine in einschlägigen Anleitungen und Wörterbüchern reglementierte und von den Trägern der jeweiligen Literatursprache bewußt als verbindlich akzeptierte historisch entstandene traditionelle Gestaltungsweise der lautlichen Seite der mündlichen literatursprachlichen Rede zu verstehen. Diese Gestaltungsweise der lautlichen Seite des literatursprachlichen Sprechens setzt vor allem die Einhaltung sowohl bestimmter Regeln, die die Artikulation der Laute und ihrer Folgen im Sprechstrom steuern, als auch des normativen Lautbestandes einzelner Wörter voraus. Sie ist wandlungsfähig und wird deshalb notwendigenfalls auf ihre Entsprechung der literatursprachlichen Sprechwirklichkeit überprüft und - wenn nötig - dieser Sprechwirklichkeit angepaßt. Eine kodifizierte AN existiert mindestens in zwei Formen, die sich voneinander durch Sphären ihrer Anwendung und den Genauigkeitsgrad in der Einhaltung normativer Vorschriften unterscheiden.

Die oben besprochenen Begriffe, die die internationale Kodifikationspraxis widerspiegeln, erlauben es in erster Linie, wie aus der erweiterten Definition der kodifizierten AN ersichtlich ist, eine Theorie der kodifizierten AN zu entwickeln. Mit Hilfe dieser Begriffe lassen sich nicht nur der synchrone Zustand und die Geschichte einzelsprachlicher kodifizierter Aussprachenormen beschreiben, sondern auch Anforderungen an Normen junger Literatursprachen, deren Festlegung noch bevorsteht.

ANMERKUNGEN

1. Vgl., z.B.: Norma i social'naja differenciacija jazyka. M., 1969; Semenjuk N.N. Norma. - In: Obščee jazykoznanie. Formy suščestvovanija, funkcii, istorija jazyka. M., 1970, s.549-596; Skvorcov L.I. Teoretičeskie osnovy kultury reči. M., 1980; Nerius D. Untersuchungen zur Herausbildung einer nationalen Norm der deutschen Literatursprache im 18. Jahrhundert. Halle, 1967.
2. Achmanova O.S. Slovar' lingvističeskich terminov. M., 1969, s. 270.
3. Ščerba L.V. Jazykovaja sistema i rečevaja dejatel'nost'. L., 1974, s. 141-146.
4. R.I. Avanesov unterschied zwischen der strengen, d.h. durchgehend normierten, und der freien, d.h. weniger normierten Abart des neutralen Aussprachestils, oder zwischen Bühnenaussprache und orthoepischem Minimum, das für jeden Träger der russischen Literatursprache absolut notwendig ist. Siehe: Avanesov R.I. Russkoje literaturnoje proiznošenie. 6-e izdanie. M., 1984, s.35-36. Eine andere Scheidung innerhalb der normativen Aussprache des Deutschen schlägt G. Meihold vor, der mit dem Begriff orthoepischer Formstufen arbeitet und neben der vollen Formstufe (gesprochene Dichtung) zwei Formstufen der deutschen Standardaussprache unterscheidet, die er als gehobene Formstufe und Formstufe des Gesprächs bezeichnet. Jede von ihnen zerfällt in zwei weitere Stufen: hohe Formstufe und gemäßigte Formstufe einerseits und gehobene Formstufe des Gesprächs und lässige Formstufe des Gesprächs andererseits. Siehe: Meinhold G. Deutsche Standardaussprache. Lautschwächungen und Formstufen. Jena, 1973 (Wissenschaftliche Beiträge der Friedrich-Schiller-Universität Jena 1973), S. 71 ff.
5. Vgl., z.B.: "Die Einheit der Literatursprache "auf dem ganzen Territorium" schließt eine gewisse Variierung ihrer Normen in einzelnen Teilen dieses Territiriums nicht aus". Siehe: Skvorcov L.I. Teoretičeskie osnovy kultury reči. S. 119. Vgl. auch: Kleine Enzyklopädie. Deutsche Sprache. Leipzig, 1983. S.383 f.

ZUR FRAGE VON STILISTISCHEN VARIANTEN IN DER DEUTSCHEN STANDARDAUSSPRACHE UND IHRER BERÜCKSICHTIGUNG BEI DER NORMKODIFIZIERUNG

EVA-MARIA KRECH

Wissenschaftsbereich Sprechwissenschaft
der Martin-Luther-Universität Halle-Wittenberg
Halle/Saale, 4020, DDR

Zusammenfassung

Die Brauchbarkeit von Normkodifizierungen der deutschen Standardaussprache hängt zunehmend davon ab, inwieweit phonostilistische Differenzierungen berücksichtigt werden. Eine systematische Untersuchung stilistisch bedingter Aussprachevarianten erfordert, unterschiedliche Grade der Artikulationspräzision in ausgewählten Kommunikationsereignissen zu ermitteln und auf ein Bedingungsgefüge zu beziehen. Es wird über eine Pilotstudie berichtet, die einen Beitrag zur Bearbeitung der Problematik darstellt.

Unter Standardaussprache wird hier die allgemein realisierte, akzeptierte und erwartete Ausspracheform auf der Ebene der Hoch- oder Literatursprache verstanden. Der Begriff bezieht sich damit nicht auf eine exklusive oder elitäre Sprechweise, schließt andererseits aber auch nicht territorial geprägte oder salopp umgangssprachliche Realisationen ein. Die Standardaussprache wird von geübten Sprechern in gelesenen, memorierten, halbfrei und frei gesprochenen Äußerungen realisiert. Ihr Anwendungsgebiet reicht von der feierlichen Rede im Großraum bis zu Gesprächen zwischen wenigen Kommunikationspartnern und beschränkt sich nicht auf öffentliche und offizielle Situationen.
Je nach dem konkreten Verwendungs- bzw. Kommunikationsbereich weist die Standardaussprache jedoch in Gestalt unterschiedlicher Artikulationspräzisionen Varianten auf, die stilistische Funktionen erfüllen. D.h., die der Literatur- oder Hochsprache eigene stilistische Differenziertheit besitzt ein Pendant in der Aussprache. Sollen Brauchbarkeit und Anwendungsbreite künftiger Aussprachenormierungen erhöht werden, dürften diese phonostilistischen Differenzierungen zu berücksichtigen sein. Untersuchungen zur Ermittlung stilistisch bedingter Aussprachevarianten schließen notwendigerweise an jene Erhebungen zum Sprechgebrauch an, auf denen die kodifizierte Norm im "Wörterbuch der deutschen Aussprache" /1/ basierte und die z.B. den r-Laut /2/, die Behauchung der Verschlußlaute /3/, den Schwa-Laut /4/, den Vokaleinsatz /5/ u.a.m. betrafen.
Diese Erhebungen hatten eine ganze Reihe von Positionen verdeutlicht, in denen satzphonetisch-intonatorisch und koartikulatorisch bedingte Varianten prinzipiell auch in der Standardaussprache gebräuchlich sind. Es handelte sich dabei im wesentlichen um Lautschwächungen, d.h. um Veränderungen, Reduktionen und Lautschwund im Konsonantismus und Vokalismus, die, vor allem infolge des im Deutschen stark zentrierend wirkenden Satzakzentes, als Ausdruck der Relaxation für die schneller gesprochenen, unbetonten Positionen kennzeichnend sind. Die Untersuchungen des Sprechgebrauchs hatten jedoch außerdem aufgedeckt, daß diese Varianten - offensichtlich abhängig von einem umfassenderen Bedingungskomplex - unterschiedlich stark ausgeprägt und unterschiedlich häufig auftreten, und zwar phonetisch insbesondere vermittelt über Sprechspannung und Sprechgeschwindigkeit.
Eine systematische Ermittlung phonostilistischer Differenzierungen erfordert somit festzustellen, unter welchen Bedingungen beim Gebrauch der Standardaussprache welche unterschiedlichen Grade der Lautschwächung eine stilistische Funktion besitzen, d.h. zur Charakterisierung konkreter Kommunikationsbereiche und zur Verdeutlichung der den Kommunikationsbereichen zugehörenden dominierenden Funktionen der sprachlichen Kommunikation beitragen.
Zunächst sei unterstellt, daß bei entsprechenden Untersuchungen die Lautung in Kommunikationsereignissen ermittelt wird, die gesellschaftlich relevant und für einen bestimmten Typ von Kommunikationsereignissen repräsentativ sind. Ein bisher keineswegs gelöstes Problem stellt jedoch die Frage dar, welche konkreten Bedingungen und Merkmale der Kommunikation im Zusammenhang mit der Wahl unterschiedlicher, stilistisch bedingter Artikulationspräzisierungen wichtig sind.

In neueren Darstellungen, bei denen es um die Untersuchung und Zuordnung /6; 7/ sowie erstmals im Rahmen einer Normkodifizierung /8/ um Empfehlungen zum Gebrauch phonostilistischer Differenzierungen geht, wurden mit den Merkmalen Vorlesen und Vortragen von künstlerischen und Sachtexten sowie freies Sprechen in unterschiedlichen Gesprächsstufen und Redeformen bereits wesentliche Kriterien berücksichtigt, die die Artikulationspräzision beeinflussen. Allerdings dürfte der Bezug auf diese Merkmale noch nicht genügen /9/. Insbesondere bleiben die Funktion der Äußerung, die Struktur des Kommunikationsereignisses und die sozialen Beziehungen zwischen den Kommunikationspartnern mehr oder weniger ausgespart. Da sich somit noch nicht auf eine Systematisierung zurückgreifen läßt, die mögliche Bedingungen, unter denen Standardaussprache realisiert werden kann, befriedigend und mit entsprechender Wichtung der Merkmale berücksichtigt, bleibt gegenwärtig nur, die jeweils konkreten Bedingungen zu erfassen und zu beschreiben, unter denen die zu untersuchende Äußerung produziert wurde. Das Ziel ist, dabei zunehmend zu einer für Normierungszwecke brauchbaren, d.h. auf einen praktikablen Umfang beschränkten, Bedingungstypologie zu gelangen, nach der sich phonostilistische Differenzierungen klassifizieren lassen.
Im folgenden werden Merkmale aus dem (umfassenderen) Bedingungsgefüge sprechsprachlicher Kommunikation angeführt, die für den Sprecher speziell bei der Wahl unterschiedlicher Artikulationspräzisionen im konkreten Kommunikationsereignis wesentlich sein können.
Dabei werden z.T. vorliegende Typikschlüssel u.ä. genutzt /10; 11; 12/:

Art der Äußerungsproduktion: z.B.
- textreproduzierendes Sprechen:
  - Lesen
  - Vortragen memorierter Äußerungen
  - halbfreies Sprechen
- freies Sprechen:
  - in verschiedenen Formen der Rede und des Gesprächs

Planung der Sprechhandlung: z.B.
- spontan - intentional

stilistische Merkmale der Textsorte: z.B.
- wissenschaftliche Texte
- nichtwissenschaftliche Sachprosa
- triviale Alltagstexte
- künstlerische Texte: z.B.
  - gebunden - ungebunden
  - 'klassisch' - 'modern'
  - großes Pathos - geringeres Pathos

dominierende Funktion der Äußerung: z.B.
- Praxisfunktion
- wissensvermittelnde Funktion: informieren
- verhaltenssteuernde u. meinungsbildende Funktion (primär rational- primär emotional wirkend): aktivieren
- phatische Funktion: Kontakt herstellen oder aufrecht erhalten
- expressive Funktion
- hedonistische Funktion künstlerischer Texte

Kommunikationsgegenstand/Thematik: z.B.
- erhaben - alltäglich
- tragisch - komisch

sprachliche Gestaltung des Textes: z.B.
- kompliziert - einfach
- abstrakt - konkret

Modalität der Themenbehandlung:
- deskriptiv - argumentativ - assoziativ

Grad der Offizialität (Geprägtheit durch einen gesellschaftlichen Auftrag)
- offiziell - nicht offiziell - privat

Grad der Öffentlichkeit
- öffentlich - halböffentlich - nicht öffentlich

Inszeniertheit des Kommunikationsereignisses bzw. der situativen Bedingungen:
- inszeniert - nicht inszeniert

Strukturmerkmale des Kommunikationsereignisses:
- einseitig - wechselseitig (Dialog / Gruppengespräch)
- direkt - indirekt
- interpersonal - medienvermittelt
- direkt und durch techn. Hilfsmittel verstärkt

räumliche Bedingungen: z.B.
- alltäglich - festlich
- Saal - mittelgroßer Raum - kleiner Raum

sozialer bzw. beruflicher Status der Kommunikationspartner: z.B.
- Berufssprecher - nicht Berufssprecher

Hierarchieverhältnis zwischen den Partnern:
- Partner gleichberechtigt
- nicht gleichberechtigt (Privilegierung / Unterordnung des Sprechers)

Bekanntheits- bzw. Vertrautheitsgrad:
- bekannt - nicht bekannt
- eng vertraut - vertraut - nicht vertraut

Größe und Art des Hörerkreises: z.B.
- Massenpublikum
- inhomogener relativ großer Hörerkreis
- strukturierte soziale Gruppe
- einzelner Hörer

Art der vom Sprecher realisierten dominierenden Ansprechhaltung, der spezifischen Gerichtetheit bzw. des kommunikativen Grundgestus: z.B.
- direkter - indirekter Ansprechmodus
- persongerichtet - sachgerichtet - ungerichtet
- ansprechen - nennen - sich aussprechen

Intensität des Hörerkontaktes:
- eng - nicht eng (locker)

Ausprägungsgrad der vom Sprecher realisierten Emotionen:
- stark - mittel - gering

Grad der muskulären Spannung entsprechend dem konkreten Ausdrucksgehalt (z.B. Zorn, Bitte, Resignation):
- starke - mittlere - geringe Spannung

In einer Pilotstudie wurde die Aussprache von 7 Berufssprechern in zwei verschiedenen Klassen von Kommunikationsereignissen auditiv untersucht /13/: beim Lesen von 3 Nachrichtentexten und von 4 Prosatexten erzählender Art. Bei diesen vom Funk gesendeten Beiträgen interessierte der Grad der Artikulationspräzision der Vokale e und i in den Artikeln 'der' und 'die'. Grundsätzlich war nicht zu erwarten, daß die volle Form, d.h. die Aussprache mit geschlossenem langen Vokal häufig zu beobachten ist, da auf die Artikel im allgemeinen nie der Wort- oder Satzakzent fällt. Sie finden sich in der Regel also stets in den spannungsloseren und auch schneller gesprochenen Passagen der Äußerung. Außerdem kommt hinzu, daß sie in ihrer Position vor einem potentiell akzentuierten Wort einem zusätzlichen Spannungsverlust unterliegen. Bei Überprüfung der Frage, ob bestimmte Ausprägungsgrade der Schwächung außerdem für bestimmte sprechstilistische Bereiche charakteristisch sind, wurde unterschieden zwischen voller Form
[de:$^{\varepsilon}$] und [di:], Kürzung des Vokals: [de$^{\varepsilon}$], [di], und zusätzlicher Veränderung der Qualität durch Öffnung des Vokals: [dɛ$^{(e)}$], [dɪ]. Weiterführende Reduktionen der Vokale kamen in dem Material nicht vor.
Folgende Hauptergebnisse ließen sich ermitteln:

(s. Tabelle)

Es zeigt sich somit vor allem:
In sämtlichen untersuchten Äußerungen herrscht die erste Reduktionsstufe (Kürzung des Vokals) vor.
Die Lautschwächung ist bei den N-Sprechern weniger häufig und weniger stark ausgeprägt als bei den P-Sprechern. Sie kann vermutlich als Merkmal für stilistische Differenzierungen im Bereich der Standardaussprache gelten. Zusätzliche Untersuchungen der Sprechgeschwindigkeit ergaben:
Die durchschnittliche Geschwindigkeitsfluktuation, d.h. der Wechsel der Sprechgeschwindigkeit von Sprecheinheit zu Sprecheinheit beträgt bei den N-Sprechern 33 Silb./Min., bei den P-Sprechern 55 Silb./Min. Damit nehmen die Vokalschwächungen eindeutig, wenn auch nicht proportional, mit der stärkeren Bewegtheit

Aussprache der Vokale e und i in den Artikeln 'der' und 'die' beim Lesen von Nachrichten (N) und Prosatexten (P)

	der		die	
Realisationsart \ Sprechergruppe	N	P	N	P
volle Realisation	41 %	25 %	38 %	24 %
Vokalkürzung	56 %	63 %	62 %	76 %
Vokalöffnung (neben Kürzung)	3 %	12 %	-	-

der Sprechweise bei den P-Sprechern zu. Bei einem Vergleich der kommunikativen Konstellationen beider Sprechergruppen wird deutlich: Unterschiede gibt es vor allem hinsichtlich der stilistischen Merkmale der Textsorten, der dominierenden Funktion der Äußerungen, der emotionalen Anteilnahme der Sprecher, des Hörerkontaktes und der Ansprechhaltung. Die N-Sprecher vermitteln Sachtexte mit der Absicht, sachbetont zu informieren. Die Ansprechhaltung ist indirekt, der Hörerkontakt locker. Eine emotionale Komponente ist nicht spürbar. Die P-Sprecher vermitteln Texte erzählender Art, die das Grenzgebiet zur Belletristik wie auch das zur Alltagsrede tangieren. Die Texte dienen der Unterhaltung und haben eine phatische Funktion. Die Sprecher simulieren eine enge Vertrautheit mit den Hörern. Meist dominiert ein direkter Ansprechmodus. Obwohl auch diese Sprecher lesen, entsteht der Eindruck einer lockeren Gesprächshaltung mit mittelstarker emotionaler Anteilnahme der Sprecher. Damit bestätigt sich wiederholt /14/, daß die bisher kaum berücksichtigten /6; 7; 8/ funktionalen und sozialen Aspekte wesentliche Bedingungsgrößen für Grade der Lautschwächung darstellen.

Literatur

/1/ Wörterbuch der deutschen Aussprache, Herausgeberkollektiv unter Leitung v. H. Krech, 1. A. Leipzig 1964
/2/ Ulbrich, H.: Instrumentalphonetisch-auditive R-Untersuchungen im Deutschen, Berlin 1972
/3/ Lotzmann, G.: Zur Aspiration der Explosivae im Deutschen, Göppingen 1975
/4/ Meinhold, G.: Die Realisation der Silben (-ən), (-əm), (-əl) in der deutschen hochgelauteten Sprache. In: ZPSK 15 (1962), H. 1/2
/5/ Krech, E.-M.: Sprechwissenschaftlich-phonetische Untersuchungen zum Gebrauch des Glottisschlageinsatzes in der allgemeinen deutschen Hochlautung, Basel/New York 1967
/6/ Meinhold, G.: Deutsche Standardaussprache, Jena 1973
/7/ Lotzmann, E.: Zur Realisierung der Standardaussprache durch Sprecherzieher. In: sprechen, Zeitschr. f. Sprechwiss., Sprechpädagogik, Sprechtherapie, Sprechkunst, Regensburg Okt. 1984
/8/ Großes Wörterbuch der deutschen Aussprache, Herausgeberkollektiv, 1. A. Leipzig 1982
/9/ Krech, E.-M.: Zu konzeptionellen Grundlagen bei der Erforschung und Kodifizierung der deutschen Standardaussprache. In: Bericht über die 14. Sprechwiss. Fachtagung, Wiss. Beiträge d. Martin-Luther-Univ. Halle-Wittenberg 1987
/10/ Steger, H., Deutrich, H., Schank, G. u. Schütz, E.: Redekonstellation, Redekonstellationstyp, Textexemplar, Textsorte im Rahmen eines Sprachverhaltensmodells. In: Gesprochene Sprache, Jahrbuch 1972, Düsseldorf 1974
/11/ Schank, G. u. Schoenthal, G.: Gesprochene Sprache, Tübingen 1976
/12/ Scharnhorst, J.: Zum Wesen des Begriffs Funktionalstil. In: ZPSK 34 (1981), H. 3
/13/ Alschner, S.: Zur Aussprache der Vokale in den Artikeln "der" und "die" in Abhängigkeit von der Geschwindigkeitsfluktuation und zwei verschiedenen Formstufen. Dipl.-Arb. Halle (unveröff.)
/14/ Krech, E.-M.: Probleme der Kodifizierung deutscher Standardaussprache. In: Beiträge zur Phonetik u. Linguistik 1987 (Festschr. f. H.-H. Wängler)

URBAN SPEECH AS A PRODUCT OF STANDARD, COLLOQUIAL AND DIALECTAL SPEECH

L.A.Verbitskaya

Dept.of General Linguistics,Leningrad State University
Leningrad, USSR, 199034

ABSTRACT

This report deals with the problem of inter- and intralanguage interference. Two types of experimental data have been analyzed: first, modern Russian pronunciation in various areas of Russia and, secondly, Russian speech of native speakers of other languages in a number of Soviet Republics.
It was found that there were similar deviations from the norm in the speech of non-native speakers of Russian,i.e.absence of palatalization and lack of i-glides in vowels, along with language specific peculiarities.
The speech of native speakers of Russian was influenced by the dialectsl, colloquial and popular features.

In our time the Russian language has become not only a tool for multinational communication within the many Republics of the Soviet Union, but rather a language used intensively in all spheres of life, as a second language in a number of republics.
Widespread modern means of communication have lead to the penetration of Russian literary language into every nook and crany of the Russian Federation, where it exerts a certain influence on the dialectal speech of many a city. And the influence of the national languages and dialects on the literary language produces deviations from the norm in the phonetic realization of phonologically essential properties of the phonetic system. The influence of the Russian literary language on dialects and national languages should be subjected to special investigation. Russian language influence on the national languages leads to the appearance of "borrowed" phonemes, alongside Russian lexical loan-words. Dialects are gradually destroyed by the effects of literary pronunciation, the sound systems are altered, although certain dialectal patterns show various degrees of stability/I, 2/.
We have observed Russian speech as it is spoken by the metropolitan population of the Russian Federation and Soviet Republics, as it is in the cities that the clash and interaction of normative and dialectal speech and colloquial speech is the sharpest.
The study of city speech may be approached in various ways. First, we may record standard Russian speech in a specific language medium and obtain a realistic picture of the language interference.Secondly, we may record the phonetic system of the Russian language in various functional conditions and define the more stable and the more mutable elements, i.e.find the weak points, elements that are subject to constant change, and the strong points that do not change.
The aim of the present investigation,conducted in the Phonetics Department of Leningrad University, is to study the functioning of the phonetic system of the Russian language from these two points of view. On the basis of comprehensive investigation of national-Russian bilingualism and intra-lingual interference, we hope to give a well-rounded description of the phonetic properties of the Russian language.
As an aid to understanding the nature of the interaction of phonetic systems, we have considered cases demonstrating the variable degrees and quality of opposition to Russian phonetic properties.
Features under consideration are the effect on Standard Russian of Russian dialects, of closely-related languages (such as, Ukrainian and Byelorussian), of cognate but not very close languages (as Latvian and Lithuanian) and of unrelated ones(Estonian, Azerbaijanian and Georgian).
For comprehensive investigation of dialectal interference, the dialects from the following cities have been reviewed: North Russian cities (Archangel,Murmansk, Vologda and Perm), Central Russian cities (Gorky, Pskov,Yaroslavl, Kuibyshev, Volgograd), South Russian cities (Smolensk, Kursk, Ryazan, Rostov-on-Don, Krasnodar), Russian cities of the Urals(Sverdlovsk, Chelyabinsk and Nizhny Tagil), and Siberian cities (Tomsk, Omsk, Novosibirsk and Krasnoyarsk).

Texts were compiled with regard to the frequency of vowels, consonants and their combinations in Standard Russian. The texts were tape recorded by groups from the cities under study of 20 to 40 native speakers representing good and poor skills in command of Standard Russian. The material was listened to by the experimentor, by a group of native subjects and then analysed experimentally. All deviations from the standard were fixed in the listeners' sheets. Findings were systematized and subjected to statistical prosessing, which revealed the most striking perceptual features and statistically significant segmental units (stressed and unstressed vowels, consonants and their combinations) and also suprasegmental features.
It is not always easy to differentiate between segmental and suprasegmental features; for instance, a lack of unstressed reduction, which must be considered segmental, leads to rhythmical alteration in the word and affects the suprasegmental construction of the utterance.
The description of the phonetic properties of dialects and national languages in their comparison with the phonetics of Standard Russian was followed by experimental analysis. Both qualitative and quantitative differences were taken into account in the comparison of phoneme inventories.
The main difficulty for the second-language learner of Russian vowels is basically the necessity of mastering an articulation that differs from the articulation of his native tongue and of accepting certain distributional rules.
As for consonants, the speakers of other languages, and even of dialects, must alter their pronunciation habits in order to produce sounds having analogies in their mother tongue, and, moreover, they must master new distinctive features, another system of oppositions, and distributional rules. It seems that the problem of mastering Russian consonants is more exacting than that of the vowels.
Then, too, both vowels and consonants are joined in syllables in speech production and its perception, so that defects of pronunciation of one group of sounds infringe on the other.
Two aspects should be distinguished in the norm, namely, orthoepy, the phoneme composition of a word, and orthophony,the manifestation of phonetic correlates of a phoneme in a word. These aspects are relatively independent. Orthophonetic distortions are possible without the disturbance of the orthoepic norm, while the phoneme structure of a word may be distorted in spite of normative use of phonemes.
The nature of Russian speech in the union republics is determined on the one hand by the characteristics of the Russian phonological system, by the specific correlation of sounds and letters and for this reason must be uniform (for example the absence of palatalizatuon, the omission of i-glides of vowels were observed in the speech of all people tested, no matter what their native language was). On the other hand, it is influenced by the native language.
The interference of the two phonetic systems is conditioned to a certain extent by the closeness of the languages. This may be true of the genetic kinship as well as the purely typological resemblance.
Moreover, the genetic affinity is not a decisive factor. Much more important are the manifestations of the phonetic system. Therefore, the interference of both languages closely akin (for instance Russian and Ukrainian) and languages that are genetically not related(Russian and Azerbaijanian) can produce sound distortions seemingly of the same type (i.e. soft sibilants, /i/ sound instead of /ы/, the appearance of /j/ in syllables where in Russian there should be a soft consonant + vowel /tja/,/mja/ and so forth. On the other hand, in some cases mistakes of this kind are completely absent when unrelated languages come in contact.
Still, in the interaction of cognate languages, the very closeness of the grammatical structure and lexical similarity encourages the use of lexemes and morphemes of the native tongue involving sound substitution even where it is not caused by phonetic difficulties.
Hence, in the study of sound interference of unrelated languages analysis of the phonetic system will be sufficient (including not only the set of phonemes but also their distribution and implementation in syllables and larger units), while in the case of cognate languages, all possible substitutions must be acounted for.
The study of Russian speech as a second language has revealed various numbers of deviations from the norm, minimal for Byelorussians, and maximal for Estonians and Georgians. These deviations are not at all alike. Some are found only in the speech of a certain language(i.e. the substitutions of sibilants by shibilants in the Russian speech of Estonians); other may be found in various languages, but their realization and phonological nature do not coincide.
Certain features and elements of the Russian phonetic system are undoubtedly difficult for speakers of other languages, who replace /ы/ with /i/ because they do not have the /ы/ sound(such as Armenian, Lithuanian, Latvian and Georgian). Or in cases where the same oppo-

sition exists but the vowels are of different qualities(i.e.Ukrainian). The nonstandard pronunciation of /ы/ is found in the Russian speech of Estonians, Moldavians and Azerbaijanians. The deviation from the standard is connected, first, with the specific articulation of the native /ы/-like sound and, second, with the incorrect articulation of the syllable itself, for example, ши – жи, the pronunciation of which is obviously influenced by spelling rules.

The speakers of all nationalities mispronounced the Russian /i/, though more rarely than /ы/. The vowel became more retracted if preceded by a partially palatalized or even non-palatalized consonant. Thus, both these traits were linked with incorrect syllable production and depended on the rules of phoneme realization in the syllable.

The commonest violation of orthophonic standards were the retention of unstressed /e/instead of /i/ when preceded by partially palatalized or non-palatalized consonants (Moldavians, Azerbaijanians, Georgians, Armenians, Ukrainians, Latvians, Lithuanians, Byelorussians and Estonians), the pronunciation of /o/ instead of /a/ and /'a/ preceded by palatalized consonants instead of /i/ when written "a" and "я" (Georgians, Ukrainians, etc.).

These mistakes are obviously caused by the different vowel distribution in the native language of the speakers, namely in the absence of vowel gradation of stressed and unstressed phonemes characteristic of Russian.

Some mistakes in vowel articulation are of orthophonic nature, i.e. the substitution of /e/ by /ɛ/, excessive vowel diphthongization and insufficient qualitative and quantitative reduction of /a/,etc.

The main difference in phonological relations between Russian and other national languages in the pronunciation of Russian consonants is the presence versus absence of consonant palatalization. In a number of languages this opposition does not occur at all(Estonian, Georgian, Armenian, Azerbaijanian). In some other national languages certain pairs are not contrasted in the same way(i.e.there is no /r'-r/opposition in Byelorussian) or palatalized consonants are produced differently than in Russian(i.e. /t', d'/ in Lithuanian). This accounts for a number of orthoepic mistakes. And here, too, there are significant differences, depending on the nature of consonants.

Voiced versus voiceless consonant opposition is observed both in Russian and other national languages investigated(except in Estonian), but in Azerbaijanian and Georgian consonants in some positions are only partially voiced. In addition, the distribution of voiced and voiceless consonants in several languages studied does not coincide with Russian. This produces both phoneme substitutions and orthophonic mistakes.

Almost all native speakers (except Lithuanians and Estonians) retain voiced consonants in the word-final position.

The largest number of deviations from Standard Russian has been registered in the Russian speech of Estonians, who do not observe such oppositions as palatalized versus non-palatalized consonants, voiced vs.voiceless and sibilant vs.shibilant. In addition, they make no contrast between fricatives and affricates.

The number of accentual traits in Russian speech of other native speakers can be listed as follows: Armenians--22, Georgians --20, Azerbaijanians--17, Moldavians --17, Latvians and Ukrainians --15, Lithuanians --14, Byelorussians--10.

Our data for this investigation have shown that phoneme infringement in word production is caused by incorrect phoneme distribution. Even native Russian speakers are guilty of such deviations from Standard Russian at times.

The majority of mistakes appear to be the result of orthophonic deviations from the Standard, i.e.insufficient palatalization, weak velarization, affricates with incorrect durational correlation of occlusive and constrictive elements, excessive diphthongization of vowels, more open or more close vowels as compared to the standard, etc.

The speech of cognate language representatives(Byelorussians or Ukrainians) also has deviations from the norm that are characteristic of colloquial speech or of popular language which indicates an insufficient knowledge of standard pronunciation.

In the speech of urban residents of large cities of the Russian Federation along with standard usage in the pronunciation of vowels and consonants we have recorded both orthphonic and orthoepic deviations. A certain set of relevant features of segmental and suprasegmental levels, a specific city pronunciation variant is to a considerable extent determined by the phonetic systems of the surrounding dialects. In Northern Russian dialects vowel peculiarities were the most striking, while in Southern Russian dialects, consonantal peculiarities stood out. The Middle Russian pronunciation variant in the main coincided with Standard Russian, only some intonation patterns being different.

The absence of a common pronunciation norm can be explained, on the one hand,by the flow of rural population bringing dialectal speech to the cities, and, on the other hand, by the comparatively recent spread of the spoken mass media, while the written literary language has a long tradition. A similar situation is found in

many European and other languages.

Older dialects have a stronger influence on speech than newer dialects. However, dialectal traits observed in the speech of city dwellers are not stable, but are found along with normative usage of vowels and consonants, and a kind of "phonetic accent" does not disturb the general perception of speech as literary, if rules of grammar and word usage are observed.

Deviations of an orthophonic nature that do not affect the phoneme composition of a word are more widespread and stable than orthoepic peculiarities(for instance, the pronunciation of a fricative/ɣ/instead of a plosive /g/).

In addition to dialectal features having a definite local occurrence, the speech of an overwhelming majority of speakers had popular or colloquial features. These were, for instance, delabialization of an unstressed /u/ (бюро́ – b'iró), nonstandard reduction of /a/, substitution of affricates /c/ and /č/ by fricatives /s/and /š/, as in "ру́чка" /rúš'ka/ and "со́лнце" /sónse/, the reduction of final /t'/ in an /s't'/ combination, such as "сла́бость" /slábas'/, etc. These same features occur in the speech of Leningraders and Muscovites, especially in the case of young people.

Substitutions of fricatives for affricates, as a most characteristic feature of colloquial speech has been described not only by specialists in Russian philology, but also by investigators of other languages, such as Slavic and Germanic.

Vowels and consonants are modified in different ways in the interaction of the native language standard,dialect and the popular language. Here the difference between interlinguistic and intralinguistic interference is strongly marked. Typical of the former is incorrect consonant articulation (a more retracted and open /e/ and /i/ which leads to distorted vowels after consonants in CV syllables, where in standard language the consonant should be palatalized, while for intralingual interference the errors in vowel pronunciation do not depend on palatalization.

The interference result is affected not only by the differences within the phonetic systems involved (phoneme differences in number, their distribution,etc.), but by how the interaction takes place.

In the interference of Russian and a native language we usually encounter an incorrect reading of the text,i.e. an error in sound-to-letter transition. The character of sound interference shows an oral approach to mastering Standard Russian.

The degree of kinship between Russian and the native language naturally asserts itself. The number of accentual traits in the Russian speech of speakers of other languages gives interesting data for further typological conclusions. In this respect languages such as Armenian and Georgian form one group, Azerbaijanian and Moldavian a second group, and Latvian and Ukrainian a third. Lithuanians in a number of accent traits occupy an intermediate position between Ukrainians and Byelorussians.

REFERENCES

/1/ L.A.Verbitskaya, "Russian Orthoepy", (in Russian), Leningrad, 1976.

/2/ L.V.Bondarko, L.A.Verbitskaya et al., "Sound System Interference"(in Russian), Leningrad, 1987.

L'OBJET ET LES FINS DE LA PHONOSTYLISTIQUE

Irina G. Torsuyeva-Leontyeva

Translator's Department
Maurice Thorez Moscow State Institute of Foreign Languages
Moscow, USSR 119034

ABSTRACT

The paper discusses the status of the phonostylistics, its reference to stylistics and other linguistic sciences, the purpose of phonostylistic studies; it also discusses the problem of phonostylistic units as secondary from the point of view of their reference to language units.

La phonostylistique est une science en formation, son statut est encore indéterminé, ses limites sont assez vagues. Ainsi, parait-il nécessaire de préciser ses rapports avec la stylistique, sa place dans la phonétique; de mettre au clair ses catégories, ses fonctions, son objet d'étude, ses unités et ses méthodes de recherches.

D'après I.Fónagy la phonostylistique a pour origine les travaux de J.Laziczius, qui distinguait les variantes combinatoires, les variantes libres non expressives, les variantes libres expressives. La phonostylistique procède d'une part de l'étude des fonctions de la langue (Bühler, Troubetzkoy, Martinet, Jakobson, Riffaterre, Léon), d'autre part des données empiriques (souvent subjectives) et des descriptions impressionnistes. Quant aux travaux stylistiques, les descriptions des formes sonores ne sont pas justifiées par l'analyse phonétique expérimentale.

Il existe actuellement plusieurs définitions de la phonostylistique vue sous l'angle de ses tâches [I,2,3,4] qu'on peut regrouper de la manière suivantes: I) la phonostylistique étudie la valeur symbolique des unités minimales de la langue; 2) elle révèle un système d'unités phonostylistiques aussi rigoureux que celui de phonèmes; 3) elle fait ressortir les valeurs potentielles de la matière sonore du message; 4) elle décrit les traits phonétiques dont l'emploi crée un effet stylistique; 5) elle étudie les variantes de discours (classes sociales, sexes, groupes d'âges et professionnels, situations de discours, etc;) 6) elle cherche à établir les règles d'encodage supplémentaire du message.

La liste de ces tâches est assez vaste, aussi pourrait-on la restreindre, compte tenu de l'existence de la dialectologie, de la sociolinguistique, de la psycholinguistique, de la phonosémantique,

dont l'objet d'étude s'avère plus ou moins clair . Ainsi, la tâche N I se rapporterait à la phonosémentique (étude du symbolisme phonétique), tandis que la tâche N 5 relèverait de la dialectologie, de la psycholinguistique et de la sociolinguistique.

Les quatre tâches restantes doivent être analysées en partant de la définition de l'objet de la phonostylistique. Cette dernière étudie les caractéristiques acoustiques du texte (message) d'où sa parenté avec la phonétique. A notre avis l'objet de l'étude purement phonostylistique peut être conçu comme le choix conscient de variantes libres des unités sonores du message. La description de ces variantes relève de la dialectologie, de la sociolinguistique, de la psycholinguistique, etc., tandis que leur étude stylistique est le propre de la phonostylistique.

Les quatre tâches qui restent se divisent ainsi en deux groupes: I) les tâches phonostylistiques; 2) les tâches à dominante stylistique (stylistico-phonétique).

Le premier groupe vise à dégager un système d'unités (phonostylèmes, d'après P.Léon) différent de celui du code phonologique (éléments segmentaux et suprasegmentaux). Par exemple, les modèles d'expression des états émotionnels; les caractéristiques acoustiques de différents types du discours (dialogue spontané, conférence, discours politique, etc.); les paramètres des registres stylistiques (niveau moyen, familier, recherché) (voir la tâche N 2).

Les tâches du deuxième groupe sont de préférence stylistique. C'est-à-dire elle visent non pas l'établissement des modèles, mais le dépouillement des contrastes phonétiques qui provoquent l'effet stylistique.

On sait que les définitions du style sont variées et contradictoires. Le style est conçu comme un choix ou bien comme un écart de la norme. Le choix se base sur la compétence linguistique et exige la connaissance des modèles, des paramètres de la norme considérée comme un étalon, un standard. Mais il est clair qu'il existe plusieurs types du discours chacun ayant sa propre norme. L'effet stylistique est également possible dans les limites de ces normes, d'où l'extrême importance de la notion du contraste.

L'étude de la distribution des contrastes phonétiques dans le message permet de révéler un encodage supplémentaire, les sèmes-clés et les sèmes potentielles.

Il s'en suit que la créativité stylistique n'est qu'une activité linguistique secondaire, qui organise le message afin d'exercer une influence sur le destinataire. On présuppose que ce dernier connaîtrait les variantes segmentales et suprasegmentales, propres aux sociolectes, jargons professionnels, patois, registres de la langue et à différents types du discours.

Dans ce cas les modèles de description (voir la tâche N 2 du premier groupe) concernent les variantes et non pas les unités. Il en découle une définition des unités phonostylistiques différente: celles-ci représentent des combinaisons de variantes libres des unités phonologiques (segmentales et suprasegmentales). Elles s'avèrent donc être des unités secondaires, le résultat de l'activité linguistique secondaire.

J.Laziczius distinguait parmi les variantes libres les variantes expressives et non expressives. Cette conception s'accordait avec la théorie stylistique des années 30 (Ch.Bally en particulier) qui étudiait les faits d'expression du language organisés du point de vue de leur contenu affectif. Si l'on reconnaît le caractère secondaire des faits phonostylistiques, il est inutile de distinguer les variantes expressives et non expressives, car une variante non expressive peut acqérir de la valeur stylistique dans un contexte. Le but de l'analyse phonostylistique est donc de révéler les valeurs potentielles des variantes libres. Dans ce cas la notion du contexte devient particulièrement importante.

L'intérêt pour la phonostylistique est justifié par la nécessité d'élaborer et d'approfondir le problème du contenu de la forme, ce qui est important notamment pour la traduction, la composition des textes publicitaires, des discours politiques, etc.

Références

I. P.R. Léon. Essais de phonostylistique. Montréal, Paris, Bruxelles. Didier, I97I.

2. I.Fónagy. Le statut de la phonostylistique. Phonética, 34, I-I8, 1977.

3. F.Carton. Introduction à la phonétique du français. Paris. Bordas, I974.

4. C.L..Van den Berghe. La phonostylistique du français. The Hague, Paris. Mouton, I976.

ATTITUDINAL SEMANTICS OF PROSODY AND ITS METALANGUAGE

YURI DUBOVSKY, GALINA YERMOLENKO

English Phonetics Dept.
Pyatigorsk State Institute of Foreign Languages
Pyatigorsk, USSR 357533

ABSTRACT

Alongside the theoretical discussion of terminology and metataxonomy problems dealing with speech prosody an approach towards setting up a correlation between a semantic label, attitude and its prosody is presented.

Prosodic features differentiating the labels that denote "friendly attitude" in English are described on the basis of the lexico-semantic and semantico-prosodic experimental data. Different degrees of descriptive power of verbal metalanguage units under study are revealed.

INTRODUCTION

Semantic description of speech prosody involves the problem of an adequate linguistic terminology and metalanguage. As it is put in /I/, a great deal of difficulties ascribed to intonation are, in fact, difficulties of inadequate metalinguistic description.

A metalanguage system is regarded as possessing a hierarchic field structure the elements of which are in hyponimic relations. It comprises a core, or a relatively closed class of generic notions (terminology) and a periphery, or a relatively open class of specific notions (metataxonomy, nomenclature).

As far as terminology is concerned, it is submitted to a formal claim made to any terminological language: terms should be neutral and monosemantic at least within the limits of a certain metalinguistic system.

Metataxonomy with labels as descriptive units has been devoted very poor attention to in contrast to terminology. There is a considerable disagreement between linguists as to what labels should be: nonverbal or verbal, artificial signs or linguistic ones, etc. The trouble is that lexical labels are not pure terms, they are borrowed from the popular speech. For this reason the majority of them are rather polysemantic than unequivocal, as terms should be. Their heavy dependence upon specific contexts creates a good deal of ambiguity and misunderstanding. Polysemy and synonymy do not seem to be the only variables that affect the choice of lexical labels,metaphorical use and evaluative colouring being the rest.

However, there is no need to reject common words as labels since any natural language possesses the metalinguistic function,i.e. it is capable of describing itself rather sufficiently. Besides, specific nonverbal metalanguages do not hold good for broad scientific descriptions.

Metalinguistic units, either verbal or non verbal, reflect two planes of description: the plane of expression and that of content, the descriptive categories being partly terms, partly labels.

The plane of expression in speech prosody can be rendered with both verbal and nonverbal descriptive units. The latter are symbolic-graphic means - prosodic transcriptions and representations. A verbal metalanguage is made up with the terminology of basic prosodic notions, i.e. units, components, structures, etc.and the metataxonomy of their specific types.

The plane of content in speech prosody is described with verbal terms and labels. Terms are used to refer to the communicative types of utterances, registers of speech, phonostyles etc., as related to the communicative and stylistic meanings of speech prosody. Semantic labels are, for the most part, of attitudinal character. They are made use of to describe the pragmatic types of utterances,of emotional and attitudinal connotations referring to the pragmatic and attitudinal aspects of speech prosody.

Any metataxonomy could be likewise characterized by a hierarchic organisation. In the attitudinal metataxonomy, for example, semantic labels fall into clusters headed by labels of a more general meaning. Related to terms as notions of higher generality labels are regarded to be specific names referring to prosodically and paralinguistically expressed emotions and attitudes.

This class of label proves to be the least systematized though the attitudinal function of intonation has been the subject of intensive study over a number of years. It may account for the fact that attitudinal labels denote psychological states

(affects, feelings). Therefore,the choice of label was habitually associated with a classification of emotions which seems more to be the province of psychology and physiology rather than linguistics. Any success in developing the attitudinal metataxonomy is hardly possible until definite suprasegmental features conveying some kind of attitude or emotion are pinned down with a specific label.
Thus, the elaboration of the metataxonomy of attitudinal prosody is highly dependent upon a set of variables: the notational system (ordinary words), meaning-relations between words (polysemy, synonymy), the essential nature of denoted referents (on the one hand, complex overlapping psychophysiological processes, on the other - suprasegmental phenomena).
From the afore said one might see that semantic labels perform two functions:
i) convey certain attitudinal meanings;
ii) point to suprasegmental means responsible for their communication.Accordingly, the study of labels could be carried out in two large spheres both involving semantic equivalence of different types.
The first sphere embraces all kinds of semantic correlations between: a) attitudes and emotions proper as between notions with their semantic volumes standing in logical relations of overlapping, inclusion, complementation and contiguity; b) attitudes and labels as between notions and lexical units; c) labels themselves as between lexical meanings (synonymy,antonymy, hyponymy).
Psychologically attitudinal labels render general emotional colour, specific attitudes and the degree of emotional intensity. These components could interact and be reflected in labels' meanings in different ways. As notions attitudes differ in several qualitative and quantitative features characterizing their psychological and physiological nature: intensity,direction, duration, way of outward expression,source and cause of origin, etc.
Intensity seems to be the most prominent feature of any affect. The information about the degrees of intensity is reflected in labels' meanings. It, thus, makes up a semantic component and could refer to either the whole meaning, or to a part of it (a single lexico-semantic variant), or even to another semantic component. Intensity could also be associated with referential properties (denotational intensity), with emotional evaluation (connotative intensity) or with both simultaneously.Since the denotational meaning of the attitudinal labels is confined to the notions of affects and feelings one might talk about the denotational (emotional) intensity reflected in their semantic structure.Intensity as a semantic feature could be revealed through the analysis of labels' dictionary explanations. This approach is possible due to the markers of intensity available in the entries. The analysis is aimed to reveal intensity-differences between labels in order to relate the data obtained to the results of prosodic analysis.
The second set of questions is connected with maintaining a three-member correlation: a semantic label - attitude - suprasegmantal means. In the frame of this correlation a semantic label refers to a bundle of distinctive prosodic features and paralinguistic phenomena carrying some kind of attitudinal meaning. This approach to the description of the attitudinal prosody seems to be more preferable as compared with the previously used technique when separate intonation patterns and pitch movements signalling emotions were explained by a great deal of ambiguous lexical means.
The hypothesis that definite suprasegmentals are fixed to specific labels has already been proposed and tested. However, the experimental evidence was concerned mainly with the auditory impressions and perceptive correlates of the attitudinal labels. The acoustic aspect of this relationship still remains undiscussed.

METHOD

In the present study our concern was a group of semantic labels referring to the prosodically manifested friendly attitude. The grouping was done on the basis of thesauruses and explanatory dictionaries. After some hesitation the total number of labels was confined to 8. The group is characterized by the field structure. At the same time it might as well be called a synonymic series. The core of it is made up by synonymous adverbs amiably, amicably, in a friendly way with the attitudinal meaning in question as a basic one. The periphery is constituted by quasisynonymous adverbs intimately, cordially, heartily, warmly, warm-heartedly with the friendly attitude being secondary or attendant. The label in a friendly way was taken as a dominant of the series.
The textual material for investigation was taken from fiction (125 samples). The samples were microsituations (of 3-5 sentences) intended to express various nuances of the friendly attitude. In these microsituations test phrases were embedded as response remarks. Test sentences (166 in total) were selected to be as colourless as possible with respect to lexics and grammatic structure. Tape-recordings of the microsituations were obtained from 3 male and 3 female native speakers (professional teachers) who simulated 8 variants of the friendly attitude corresponding to 8 labels discussed. The auditory analysis was arranged in 3 series. In the first series 5 trained listeners were to assess whether the tape-recordings sounded natural and had any

connection with the friendly attitude.The satisfactory examples were chosen (140 in total) to be later subjected to the instrumental analysis.
In the 2-nd series 10 linguistically naive native speakers were presented the tape-recordings of the satisfactory test phrases isolated from the contexts in advance. The hearers identified the attitudinal variants with most appropriate and accurate descriptive terms while the tape-recordings were played. To do the task the informants were provided with lists of 17 labels (8 test labels and 9 additional ones). They were also permitted to resort to any other descriptive categories they wished. Listening was repeated as the informants wished.
Some time later the hearers were played the whole microsituation recorded to do the same task. This was thought reasonable to indicate the effects of contexts on attitude recognition.
In the 3-rd series 6 informants (Russian teachers of English) who had training in phonetics made prosodic transcriptions of the recorded test phrases. They made use of the set up type of symbolic-graphic marking to describe prosodic features.
The instrumental analysis dealt with the acoustic correlates of prosodic features, which were interpreted as it is suggested in /2/.

RESULTS

The detailed lexico-semantic analysis has shown that friendly attitude is present in the dictionary explanations of all labels discussed. The one exception to this observation occurs in the label warm-heartedly. There is only one explanatory dictionary /3/ where the attitudinal meaning was found to be directly mentioned in the entry of this label. In the rest of the dictionaries it is implied by the near-synonyms cordially and heartily.
Observations of the other implicit markers of the attitudinal intensity exhibited differences between the labels in this respect. Attitudinal intensity was the greatest in the core of the group and the least - in the periphery of it. One can talk about the many-sided nature of the attitudinal intensity displayed by the labels: general (in a friendly way), specifying (amiably, amicably), derivative (intimately), additional or attendant (cordially, heartily, warmly, warm-heartedly).
The analysis made it possible to distinguish 3 degrees of friendly attitude intensity exhibited by the labels under consideration: high (in a friendly way, amiably, amicably), moderate (intimately) and low (cordially, heartily, warmly, warm-heartedly).
As far as emotional intensity of friendly attitude is concerned significantly different results were obtained. Heartily and warmly available in the entries of the other labels can be regarded as indirect markers of high emotional intensity, either general (for warm-heartedly), or related to friendly attitude (for cordially). In this sense, labels heartily and warmly perform the function of internal intensives.
External intensives have not been revealed in the dictionary explanations of the 'core' labels and in intimately. This testifies to the moderate nature, with regard to emotional intensity, of the attitude in question displayed by them. We also failed to detect internal intensives concerning friendly attitude in the explanations of these labels. Some of the internal intensives found out for the 'core' labels characterize positive evaluation rather than friendly attitude.
In contrast, internal intensives revealed for the 'periphery' group contribute to increasing the degree of emotional intensity.
As a result 3 degrees of emotional intensity pertinent to friendly attitude were stated: high (heartily, warmly, cordially), moderate (intimately,warm-heartedly) and low (in a friendly way, amiably, amicably).
We can easily deduce from what is said above:attitudinal intensity of the meaning 'friendliness to smb.' is reciprocal to its emotional intensity.
In listening experiments involving isolated test phrases label identification was greatest for heartily (90% accuracy) and somewhat less for in a friendly way, cordially (the percentage of correct identifications ranged from 70 to 60% of the instances). These were closely followed by intimately, amiably (60-50% of cases). For amicably, warmly, warm-heartedly identifications were considerably reduced (the recognition score was no higher than 10%). With warm-heartedly identifications were completely random.
As far as labels amicably and warm-heartedly are concerned, their poor identification may be accounted for several reasons. First, it is probably caused by either too specific (amicably) or too amorphous(warm-heartedly) lexical meanings they possess. Second, this may be due to the absence of descriptive power, devoid of any specification with respect to suprasegmental means. If the latter is true,these labels are useless as descriptive terms. However, more research is needed to confirm their descriptive inability.
Frequency of use could give rather valuable additional data as to the descriptive status of labels. There is a tendency for easily identifiable labels to be frequently used as descriptive terms of other attitudinal variants, while the reverse is true for hardly recognizable labels. However, there is no regular interdependence between correct identifications

and frequent use. On the whole, labels of more general meaning of friendliness tend to be frequently used to refer to other attitudinal variants (in a friendly way, intimately, cordially). This is not true for easily recognizable heartily. The matter is that heartily is very often used to label general emotional colour. That is why listeners often ascribe descriptive terms emphatically or impatiently to this attitudinal variant.

Label identification of test phrases pronounced within the context was, to some extent, negligible. This finding is in agreement with the results of label identification experiments done by D.Crystal /1/. Easily identifiable labels tend to have high recognition scores in both cases (in a friendly way, amiably, cordially, heartily). Amicably and warm-heartedly proved to have similar identification (10-15% of correct instances). The exception could be made for warmly. The influence of context was rather strong in this case; correct identification rose to 40%.

These observations suggest that labels could be used as terms out of context. However, the statement requires experimental confirmation since attitudinal variants rarely were ascribed a single label.

Analysing the prosodic features of the attitudinal variants under study we have obtained a) constantly overlapping, b) variationally overlapping characteristics and c) distinctive features by which a certain label differs markedly from the rest. The latter two are briefly outlined below.

Amiably - b) no instances of high initial fundamental frequency (F_0) levels, few instances of mid-narrow F_0 ranges in pre-heads, relatively low F_0 peak values; absence of medium-zone mean syllable duration; low minimum-zone intensity of unstressed syllables;
c) low decreased-zone intensity of unstressed syllables.

Amicably - b) relatively high mean values of mid-wide F_0 range in terminal tones, few instances of mid-narrow F_0 range in the utterance;
c) high mid-narrow F_0 range average values in terminal tones.

Friendly - b) moderate reccurrency of wide F_0 range in the utterance; high upper limit of high-wide and full F_0 registers in terminal tones and heads;
c) low F_0 medium range mean values in terminal tones; low minimum-zone average intensity of unstressed syllables, high medium-zone average intensity of unstressed syllables in preheads.

Intimately - b) frequent wide F_0 range in the prosodic structure; low medium-zone intensity of stressed syllables;
c) high mean values of mid-wide F_0 range in terminal tones and heads, low F_0 minimum values of nuclear syllable followed by a post-nuclear syllable; high decreased-zone mean syllable duration.

Cordially - b) high frequency of mid-narrow and mid-wide F_0 registers in prosodic structures; low decreased- and medium-zone intensity of unstressed syllables;
c) low mean values of mid-narrow F_0 range in pre-heads.

Heartily - b) instances of extra-high F_0 final level, no cases of mid-narrow and mid-wide F_0 registers in pre-heads; increased upper limits of wide F_0 register in prosodic structures, increased F_0 peak values; high F_0 minimum values; high minimum-zone mean intensity values of unstressed syllables;
c) increased upper limits of high-wide F_0 register in terminal tones; high F_0 peak values in nuclear syllables; high decreased-zone mean syllable duration.

Warmly - b) no instances of mid-wide and narrow F_0 range in terminal tones, no instances of narrow and mid-wide F_0 range in prosodic structures as compared to the 'core' labels and labels heartily, warm-heartedly; relatively high mean values of mid-narrow and wide F_0 range, particularly in comparison to other 'periphery' labels;
c) high minimum-zone shortest duration.

Warm-heartedly - b) mid-high, low, extra-high and mid-low (in decreasing rank of frequency) F_0 initial levels; high average values of minimum-zone intensity of unstressed syllables;
c) high-narrow F_0 range mean values in pre-heads.

As the results of the experiment show the most marked and functionally loaded are heartily, in a friendly way, intimately. They are most readily identified by the listeners too. The fact that certain prosodic structures are associated with a certain label makes it possible to speak about such labels as having a strong degree of descriptive power. On the contrary, labels with only some distinctive prosodic features 'attached' to them are of very little metalinguistic help (amicably, warmly, warm-heartedly) or of no use at all (such as 'amicably').

REFERENCES

/1/ D.Crystal, "Prosodic Systems and Intonation in English", London: Cambridge Univer.Press, 1969.

/2/ Ю.А.Дубовский, "Анализ интонации устного текста и его составляющих", Минск, 1978.

/3/ Longman Dictionary of Contemporary English, Longman Group Ltd., Harlow, 1978.

ATTITUDINAL AND DIALECTAL VARIATION IN INTONATION
High tone displacement and the role of the distortional component in Autosegmental theory

ANTONY HIND

Département de Recherches Linguistiques
Université Paris VII
Paris, France 75005

ABSTRACT

In this paper I show how the Autosegmental theory associated with the "distortion as a secondary message" theory of Fonagy [1] can contribute to our understanding of attitudinal variants as significant distortions from expected target realizations (or "phonotypes") of underlying segmental phonological contours (or melodemes) [2], and I discuss the relation between attitudinal and dialectal variation.

INTRODUCTION

According to this new theory :

1) the phonological contour is part of an ideophonic coding system : the juxtaposition of tonal segments defined with binary features, allows the speaker to express his attitude in relation to his assertive choices - i.e. modality.

2) modulation of the basic contour is interpreted as a significant distortion from the target contour; which is due either to :

a) non application of the obligatory rules of the intonation component (digital coding) ;

or, b) a difference in utterance structure, introduced by the application of some optional rule - i.e. modulation.

Here, I will first concentrate, in I and II, on cases of "abnormal" pitch prominence in variant R.P. contours (i.e. where pitch promininence does not correspond to underlying accent prominence as predicted by the obligatory tone/text association rules).

Finally, in III, I raise the question of whether, a) intonation variants between R.P. and Scottish English (Rise-fall replacing Fall) could be due to a special case of 2b, (tone displacement, an optional rule of R.P. could be an obligatory rule for Scottish) [4] [5] ; or b) whether, (as for some French speakers of English), different principles of tone text alignment could result in the simple Falling contour being associated twice to the same word, giving a complex Rise-falling contour [3].

I - INTENSE VARIANTS

Phonetic and semantic date brought to light during psycho-phonetic tests brought me to distinguish two contrasting phonological contours underlying Falling phonetic contours. These as in (1) are roughly equivalent to the traditional Falling and Rise-falling contours.

```
       *--          --*
(1)   H L +L (F), and L H +L (R-F),
```

By postulating an [H L +L] Falling contour contrasting with the [L H +L] Rise-falling contour, I can explain :

1) the difference of function of these two types of contours as described by Sag and Liberman [6] - which can't be accounted for in Liberman [7] - and, more important for our present discussion ;

2) the symmetric mirror image form of the neutral and intense variants of the two contours.

The intense variant of the Falling contour (Leben [8], can now be expressed in terms of a high tone escaping the effects of an obligatory Down-drift process [2].

For example, in (2.a) below, the neutral Falling contour is obtained first by the association of the contour to the text (with the "high nuclear tone" associated, by *copy*, to all the pre-nuclear accented syllables) ; this is, then, followed by the application of Down-drift, which lowers each successive high tone by one degree, giving a Falling stepping head, as in the following example :

```
          +                        +         *
(2a)  I  followed  her to a  tiny  apartment #
          H                        H^-1      H^-2 L L
```

According to Leben [8], p.85 : "*In English failure to lower the peak of the nuclear syllable is often a signal of amazement or concern*". Now in my theory, Down-drift can be formulated so as to be blocked before a certain type of pause, or rupture, (noted here with a square pause sign) ; when this is placed, optionally, before the nuclear syllable as in (2.b) :

```
          +                        +           *
(2b)  I  followed  her to a  tiny  #  apartment #
          H                        H^-1        H^-0 L  L
```

The association of this pause before the word containing a nuclear syllable results in an abnormally high nuclear syllable which is very probably judged as a "distortion" in relation to the phonotype of (2.a), (i.e. +2 degrees) which could either express a tense attitude on the part of the speaker (amazement, surprise, etc. via increased vocal tension) or simply be a means for focussing the attention of the addressee on the word containing the nuclear syllable.

In R.P., the Rise-falling contour, however, has a rising stepping head : the low tones associated to the pre-nuclear stressed syllables by copy from the low initial tone of the contour are affected by Up-drift as in (3.a).

This process of Up-drift can also be blocked before the nuclear syllable for the purpose of expressiveness or contrast ; but in this case it gives an abnormally **low** departure of the nuclear syllable as in (3.b). Only the existence of a low initial tone, in the underlying melody, can explain this abnormally low departure (-2 degrees).

```
        +          +                 *
(3a) What  a  marvellous old  steam  engine #
                                 -¯*
     L          L+1            L+2 H          L

        +          +                  *
(3b) What  a  marvellous old  #  steam engine #
                                 ¯-*
     L          L+1            L+0 H          L
```

The fact of postulating the [L H +L] contour, underlying **Rise-falling** intonation, allows me to derive all the interpretations of these contours, pointed out by Liberman and Sag [9] or myself, through an ideophonic interpretation of its components : the nuclear contour in [LH] would support interpretations, compatible with the notion of increased vocal cord tension : "appeal to the addressee", "astonishment", "putting the addressee in question" etc.. However, it is clear that the **Rise-falling** and **Falling** contours have in common their non-open character, and this would be present in both contours, on the level of the final floating low tone, that they have in common.
However, it should be noted, that the negative or positive nature of speaker's judgment (admiring surprise, or reproving astonishment) depends more on the context of the utterance, paralinguistic features etc. rather than the contour itself.
The same low departure for the nuclear syllable is also heard in ironical realizations of the Rise-falling contour as in (4).

```
     +          +        +          *
(4) She says her husband's name is # Pam e la #
                                     *
    L          L+1       L+2     L+0 H    L
```

But, here, there is also a tendency to displace the nuclear tone on to a following post-nuclear syllable by High tone displacement, creating a further "surprise" effect. Speaker intrusion is manifested by the disturbed accent/pitch structure.
Roughly then, the more the contour differs from that predicted by the obligatory rules, the greater the speaker intrusion in his text. Note that the presence of the initial low tone, explains why the High tone can not move to the left. As predicted by this analysis, we also find the mirror image of this contour with falling intonation, where the nuclear High tone is retracted on to a preceding weak syllable (the presence of post-nuclear low tone prevents its realignment to the right) :

```
                 -        *
(5a) He  has  opened  the  door  #
     M                 H    L    L
           +    +          *
(5b) He  may  open   the  door  #
     M    H     H     H    L
```

In (5a) the nuclear height of "the" is increased relatively by deaccenting the pre-nuclear syllables. The unexpected high tone on "the" is particularly apt for expressing the speaker's "shocked surprise". While in (5b), the combination of level pre-nuclear contour (non-application of Down-drift) - giving an "abnormally" stylized contour suitable for stereotyped performative exchange (cf. [2], [10], [11]) - and, unexpected High tone on "the" - speaker intrusion in the text, expressing surprise or irony - apparently can be used to evoke an ironic effect, expressing "mocking authoritative contempt" (cf. [2], [8]).

II - STEREOTYPED EXCHANGE, AND PITCH/ACCENT MISALIGNMENT

Some cases of tone prominence of non-accented syllables even in R.P. do not give rise to "surprised/emphatic contours". Indeed, in certain rhythmic realizations - where the metrical grid imposes strict alternating strong/weak patterns - a weak, or even reduced, vowel can be aligned with a strong metrical position, and have a high or low tone associated to it ; but in this case, the position of the tone is totally predictable from the rhythm rules, and as such, is suitable only for passing a message with low pragmatic information content and is generally part of a stereotyped exchange.
For example, in a highly rhythmic form of the Rise-falling contour, an utterance such as "Canada is green" is aligned with a trochaic (alternating strong/weak) grid, and all pre-nuclear strong grid positions receive a low tone,

```
      +        +         *
(6) Cana§ da is    § green #
     L  H  L        H LH   L
```

(where "§" is a rhythmic juncture)
even though the Metrical grid does not coincide with the underlying rhythmic structure in which the last syllable of Canada is weak and the vowel reduced.
This realization while being very emphatic, is nevertheless highly predictable in the position of the tones, and is most probably found as a desperate repetition "just for the form" (I've told you before, but I can see I will have to tell you again). It can never be used to express alarm or surprise from the very nature of its predictability.
The claim is thus that there would be a close relation between predictability of contour and the conventional nature of the communication.
The distortion theory, then, makes the prediction that misalignment of pitch and accent structures will be used to vehicle attitudinal variations or modulations within the same dialect ([2], [5]), and that the relation between contour and accentual structure must therefore be rule governed. Furthermore, it correctly predicts that speakers of dialects which differ in their contour accent/structure association will tend to be received by each other as deviating significantly on the attitudinal level (i.e. their utterances will seem to continually vehicle secondary messages of the same type, when judged from the standpoint of the speaker of the other dialect). This may well at least partly explain certain stereotype characteristics which one linguistic group tends to attribute to another.

III - DIALECTAL AND INTERLANGUAGE VARIATION

When dealing with dialectal variation, it is not always easy to know whether we are dealing with differences of accentuation structure, phonological contour, or surface differences due to "displaced nuclear tones".
Listening to a recording of a Scottish regional speaker. I became aware of what at first seemed to be accentuation variants as follows :

```
(7a)  Exciseman,        Smugglers
       L  H  L            L  H L
```

(though the initial syllable in each case sounded heavier than the the syllable on which the rise occurred) ; whereas normal R.P. pronunciation with Falling intonation would give :

```
(7b)  Exciseman,        Smugglers
       H  L               H   L
```

However this put me in mind of the way French students tend to repeat certain stress patterns produced by an R.P. English speaker with falling intonation even when they know the stress rules of English.
The facts are as follows : when the student repeats a word such as "development", frequently the intonation peak is on "op" and not on the stressed syllable 'vel'. At first I thought these errors were also due to an error in stress positioning but when on many occasions the students making the "error" claimed that their intention was to stress the target syllable, I realized it was no doubt a problem of contour. My first analysis of the facts [3] was as follows : the student would in fact have associated two Falling contours to the same word giving the pattern :

```
(8)       *                    *
     de vel opment         fath er
     H   L  H L            H L    HL
```

Roughly, the first contour would be associated to the stressed syllable according to the principles of English (though, the form of the association would follow the rules of French, i.e. One tone per syllable and one syllable per tone) ; however a second contour would then be associated to the end of the word entirely according to the rules for associating French contours). The resulting contour would be due to the application of two different sets of language principles to the same word. The student would still, according to this hypothesis, be trying to reproduce a Falling contour, the unfortunate Rise-falling contour would be a phonetic rather than a phonological fact. I presented this hypothesis in a communication in 1984 where I suggested (very tentatively) that a similar Celtic/English bilingualisme might have been the origin of the Rise-falling contour of Scottish English which often seems to replace the neutral Falling contour of R.P..
Since then D.R. Ladd made me aware of a rather different analysis in which he suggested that High tone delay features could explain this sort of case. In his analysis the Rise-falling contour is a variant of the Falling contour with optional delayed High nuclear tone, and dialectal variants could have obligatory rather than optionally delayed High nuclear tones ([4], [5]). With my analysis of R.P., presence of the post-nuclear low tone in the Falling contour, blocks realignment to the right. However, this does not necessarily preclude the possibility that Scottish speakers, as well as French speakers of English, might be using a simpler Falling contour as follows : [MHL]. In that case High tone delay would also be a possible analysis within my theory.
The question remains whether it is possible to prove either analysis and I think it might be. For example, on a longer word such as "terrifying" if the peak is on the post nuclear syllable and not on the prefinal syllable it could be an argument in favour of the hypothesis of delayed High tone. I have not as yet been able to follow up this problematic as far as Scottish speakers are concerned but I have been able to increase the number of examples obtained with French students speaking English.
The results in (9a) (though not those in (9b)) at first examination seem to suggest that delayed High tone is the valid explanation as the peak is clearly post-nuclear.

```
        *                    *
(9a) terr i  fy  ing      plan  et  ar  y
      L   H   H   L         L    H   H   L
            *             *                  *
(9b) economically or economically cy ber net ic
      L H     H L      L H  H L  L H   H   L
```

However when you point out to a speaker (using a diagram) that the English contour for (9a) goes down on "ter" and "plan" but stays down over the whole word, the French speaker very generally stays down over the first two syllables, rising however on the pre-final syllable as follows :

```
(10)   *                    *
     terr i fy ing       plan et ar y
      L   L H  L          L   L  H L
```

This tends to imply that the initial low tone here can not simply result from High tone displacement, and that it results from a partly independent phonological choice.
I suggest that these realizations do in fact result from the association of two contours but that the form of the initial contour is at least partly determined by the rules of French.
In certain emphatic forms of French intonation the initial syllable of a word is lowered and yet a normal Falling contour effects the end of the word. This complex contour seems to be made up of an LH associated to the beginning of the word and a final HL associated to the end of the word (the tones being assigned according to the biunique principle: one tone/one syllable). This contour on the word "terrifiant" would give the following result :

```
(11)      terr  if  i  ant
           L    H   H   L
```

This is very close to the realization given above on terrifying.
Now, if this is the correct analysis we understand how a French speaker might try to extend the association of the initial low tone over the following syllable normally aligned with the High tone as follows :

```
(12)  terr i fy ing  →  terr i fy ing
       L   H H L         L   L H  L
```

It would be difficult to see how High tone delay would afford an explanation of these cases.
Now, in (8) above, I suggest that the French students would have associated a slight deviant from this contour (with the LH associated to the second rather than initial syllable) on "development" simply because it is the nearest thing to the English contour that French can offer (i.e. with a pitch movement anywhere else than at the end of the word).
More significantly, my recordings also show that French students not only use this contour when the nucleus is near the beginning of the word but also where there is an initial secondary stress as shown

above in (9b). This explains why they tend to neutralize the difference between a word like "cybernetic" /2010/ and a word like "terrifying" /1000/ as follows :

```
      +   *          *
cybernetic      terrifying
 L H  H L        L   H  H L
```

This fact is easily explained within the present analysis but not within the hypothesis of delayed High tone.

It must be noted however that this result shows that a strong universal gestuel theory (as criticized in Ladd [4]) is quite untenable. In this example, drawn from French student's attempts at reproducing English contours, it is shown that contour form is far more important than contour meaning. Attempts at reproducing the contour shape are more important than the meaning conveyed (motivated or otherwise). An emphatic French contour is substituted for a neutral English contour simply because it is the contour with the closest phonetic shape.

References

[1] I. Fonagy, "La vive voix, essais de psychophonétique", Payot, 1983.

[2] A. Hind, "Phonosyntaxe : Place et Fonction de l'Intonation dans une Grammaire". Thèse de Doctorat d'Etat, non-published, Université Paris VII, 1986.

[3] A. Hind, "Research on English intonation in an autosegmental framework", C.E.L.D.A. : le Suprasegmental, Université Paris Nord, Villetaneuse, avril 1984.

[4] D.R. Ladd, "On Intonational Universals", The Cognitive Representation of speech, T. Myers et al. eds., North Holland Publishing Company, 1981.

[5] C. Gussenhoven, "On the grammar and semantics of sentence accents", Dordrecht : Foris, 1984.

[6] I. Sag, and M. Liberman, "The Intonational disambiguation of indirect speech acts", C.L.S. 11, 1975.

[7] M. Liberman, "The intonational system of English", Indiana University Linguistics Club, 1978.

[8] W. Leben, "The Tones in English intonation", Linguistic analysis 2, p. 69-107, 1976.

[9] M. Liberman and I. Sag, "Prosodic form and discourse function", C.L.S. 10 : 416-27, 1974.

[10] D.R. Ladd, "The Structure of intonation meaning", Indiana University, 1980.

[11] I. Fonagy, § al., "Clichés mélodiques", Societas Linguistica Europea, p. 273-303,1983.

DIE ROLLE DER TONHÖHE IN DER EMPHASE AM BEISPIEL DES KATALANISCHEN

JAUME TIO

Laboratori de Fonètica "Pere Barnils"
Estudi General de Lleida/Universitat de Barcelona

ZUSAMMENFASSUNG

Es ist bekannt, daß man sich im Alltagsleben sehr oft von der Emphase bedient, um jene Empfindungen auszudrücken, die ihm am tiefsten Herzen liegen. So wie man im Katalanischen feststellen kann, sind solche emphatische Sätze von den normalen Aussagesätzen syntaktisch gesehen oft nicht zu unterscheiden. Ziel unserer Untersuchung ist es, die Rolle zu enthüllen, die die Tonhöhe bei der Anerkennung emphatischer Sätze seitens des Hörers spielt. Bei diesem Versuch wird man feststellen, daß die Emphase mit bestimmten Tonhöheschwankungen verbunden ist, die alleine imstande sind, uns zu zeigen, wann der Satz eine Emphase enthält und wann nicht. Um dieses zu veranschaulichen, haben wir verschiedene Sätze von mehreren Informanten aufgenommen, sie nach der Tonhöhe ihrer Elemente klassifiziert und unsere Schlußfolgerungen durch einen Wahrnehmungstest bestätigen lassen.

EINFÜHRUNG

Die Untersuchung wurde im *Laboratori de Fonètica "Pere Barnils"* durchgeführt. Grundsätzlich haben wir mit dem Elektroglottograph (F-J Electronics ApS, Modell EG 830), einem oszilloskopischen Schirm (F-J Electronics ApS, Modell CD 1300) und einem X-Y Register (Hewlett-Packard, Modell 7010B) gearbeitet, mit denen wir die Intonationskurven von insgesamt 75 Sätzen (15 Sätze x 5 Informanten) gewonnen haben. Sie wurden gleichzeitig auf Tonband aufgenommen (UHER 4400 Report Monitor) und auf Papier gedruckt. Teilweise haben wir auch mit dem Visi-Pitch (KAY Elemetrics, Modell 6087) und dem Digital-Sonagraf (KAY Elemetrics, Modell 7800) gearbeitet, um mangelhafte Information nachzuholen (z.B., um festzustellen, wo oder in welcher Silbe die gesuchte Tonhöhe lag).

Ziel der Untersuchung war, die Rolle zu enthüllen, die die Tonhöhe bei der Produktion der Emphase spielt. Wir sind von der Hypothese ausgegangen, daß die Emphase sich wohl durch verschiedene Tonhöheschwankungen charakterisieren würde, abgesehen von anderen suprasegmentalen Elementen wie Tempo und Intensität.

CORPUS

Wir haben fünf verschiedene Satztypen ausgewählt und für jeden Typ drei Sätze zur Untersuchung gestellt:

A1 *Els llibres, els va dur el pare.(Les llibretes, la mare)*
A2 *Les pomes, se les va menjar el nen.(Les peres, la nena)*
A3 *Els arbres, els va matar el fred.(Les plantes, la sequera)*
B1 *Els llibres va dur el pare.(No pas les llibretes)*
B2 *Les pomes es va menjar el nen.(No pas les peres)*
B3 *Els arbres va matar el fred.(No pas les plantes)*
C1 *Els llibres va dur el pare?(Em pensava que les llibretes)*
C2 *Les pomes es va menjar el nen?(Em pensava que les peres)*
C3 *Els arbres va matar el fred?(Em pensava que les plantes)*
D1 *El pare va dur els llibres.(La mare, les llibretes)*
D2 *El nen es va menjar les pomes.(La nena, les peres)*
D3 *El fred va matar els arbres.(La sequera, les plantes)*
E1 *El pare va dur els llibres.(No pas la mare)*
E2 *El nen es va menjar les pomes.(No pas la nena)*
E3 *El fred va matar els arbres.(No pas la sequera)*

Die A-Sätze sind dadurch charakterisiert, daß die beiden Teile des Satzgefüges eine Gegenüberstellung darstellen, in denen das direkte Objekt thematisiert und infolgedessen vorangestellt wird; das direkte Objekt wird dann durch das jeweilige Pronomen ersetzt. In den B- und E-Sätzen ist auch eine Gegenüberstellung zwischen beiden Satzgefügen zu erkennen; der Unterschied liegt nur daran, daß das direkte Objekt in den B-Sätzen umgestellt wird. Schließlich sind die D-Sätze Aussagesätze, deren Satzgefüge verschiedene Tatsachen aufzählen, und die C-Sätze Fragesätze, in denen auch eine Gegenüberstellung zwischen dem Gedachten und dem Wirklichen vorliegt.

INFORMANTEN

Die Informanten waren fünf Studenten der Philosophischen Fakultät, 17 bis 22 Jahre alt, männlich; sie haben ihr ganzes Leben im westkatalanischen Gebiet verbracht. Geburtsort und Wohnort der Informanten sowie ihrer Eltern liegen wiederum im westkatalanischen Gebiet, bis auf die der Mutter des fünften Informanten, die in Ostkatalonien geboren wurde, seit langem aber im Westen wohnhaft ist.

WAHRNEHMUNGSTEST

Der Test enthielt nur zwei Fragen, zu denen es drei Antwortsmöglichkeiten gab.

1.Was für einen Satz ist das?
- *0. Es ist zweifelhaft/Ich weiß es nicht/...*
- *1. Aussagesatz*
- *2. Fragesatz*

2. Wie ist dieser Satz?
- *0. Es ist zweifelhaft/Ich weiß es nicht/...*
- *1. Neutral*
- *2. Emphatisch*

Der Test wurde von 18 Studenten aus der Philosophischen Fakultät beantwortet, die alle zum westkatalanischen Dialekt gehören.

ERGEBNISSE

Aus den gewonnenen Intonationskurven kann man schließen, daß wir vor vier Grundtypen stehen. In all den Typen fängt die Kurve mit einer kleinen Abnahme der Tonhöhe an, die gleich danach den höchsten Punkt erreicht, abgesehen von den Fragesätzen, wo manchmal der höchste Punkt am Satzende liegt. Die vier Grundtypen unterscheiden sich wie folgt:

1. Typ. Nach dem ersten Gipfelpunkt nimmt die Tonhöhe ohne bedeutende Schwankungen allmählich ab.

2. Typ. Nach dem ersten Gipfelpunkt nimmt die Tonhöhe zuerst mal stark, dann schwächer ab, so daß eine Beugung in der fallenden Kurve festzustellen ist.

3. Typ. Nach dem ersten Gipfelpunkt nimmt die Tonhöhe auch zuerst mal ab, um dann aber wieder bis zu einem zweiten sekundären Gipfelpunkt zu steigen, bevor sie wieder abnimmt.

4. Typ. Nach dem ersten Gipfelpunkt nimmt die Tonhöhe zuerst mal ab, bis sie einen Tiefpunkt erreicht, aus dem die Kurve steigend zum Satzende geht.

Aus den elektroglottographischen Aufnahmen haben wir folgende Ergebnisse gewonnen. Die Zahlen entsprechen den Hz-Messungen in folgenden Punkten der Intonationskurve: 1=Anfangspunkt, 2=erster Tiefpunkt, 3=erster Gipfelpunkt, 4=zweiter Tiefpunkt oder Beugung, 5=zweiter Gipfelpunkt und 6=Endpunkt.

Satz	Hz/1	Hz/2	Hz/3	Hz/4	Hz/5	Hz/6
1-A1	115	100	190	120	150	60
1-A2	140	110	200	135	175	75
1-A3	130	105	190	95	150	60
1-B1	150	120	200	115	140	70
1-B2	155	120	205	120	155	70
1-B3	145	130	185	95	125	80
1-C1	125	110	210	115	---	205
1-C2	140	110	215	110	---	225
1-C3	130	100	210	80	---	175
1-D1	115	80	180	---	---	65
1-D2	150	130	185	135	165	70
1-D3	150	100	185	135	150	60
1-E1	155	140	180	---	---	65
1-E2	150	130	190	135	175	60
1-E3	150	135	200	105	135	60
2-A1	165	135	250	140	---	235
2-A2	145	110	250	145	175	85
2-A3	155	135	250	135	---	200
2-B1	140	125	235	130	---	240
2-B2	150	90	215	---	---	95
2-B3	160	140	235	125	---	210
2-C1	170	150	260	150	---	265
2-C2	175	135	250	145	---	225
2-C3	150	125	250	130	---	225
2-D1	155	115	215	---	---	90
2-D2	165	125	220	---	---	95
2-D3	175	145	220	---	---	100
2-E1	155	120	220	---	---	105
2-E2	155	135	225	---	---	80
2-E3	150	130	210	---	---	100
3-A1	105	85	180	90	115	40
3-A2	115	85	170	110	125	75
3-A3	100	90	165	105	120	50
3-B1	110	100	165	60	90	50
3-B2	110	85	145	70	---	50
3-B3	110	95	150	75	95	40
3-C1	110	100	185	60	---	190
3-C2	105	85	180	85	---	185
3-C3	105	100	185	75	---	165
3-D1	110	90	160	---	---	40
3-D2	140	100	180	---	---	55
3-D3	100	80	165	---	---	65
3-E1	115	75	170	---	---	65
3-E2	95	80	165	---	---	50
3-E3	100	75	155	---	---	40
4-A1	145	110	185	125	165	120
4-A2	130	120	195	145	---	55
4-A3	135	125	180	---	---	140
4-B1	140	115	180	115	---	80
4-B2	140	120	180	110	---	75
4-B3	135	110	190	---	---	125
4-C1	155	120	165	75	---	200
4-C2	135	100	165	65	---	205
4-C3	150	115	155	100	---	135
4-D1	130	110	180	150	160	85
4-D2	165	150	190	150	165	115
4-D3	130	110	190	140	150	125
4-E1	140	115	185	145	160	105
4-E2	125	100	160	115	155	110
4-E3	150	110	185	135	150	75
5-A1	180	155	255	170	200	130
5-A2	190	145	295	190	220	135
5-A3	180	170	275	170	215	130
5-B1	215	175	280	155	180	140
5-B2	190	155	255	160	175	135
5-B3	170	150	215	155	---	120
5-C1	200	175	260	165	---	260
5-C2	185	155	275	155	---	300
5-C3	205	165	285	165	---	285
5-D1	185	135	250	---	---	135
5-D2	210	200	245	190	230	260
5-D3	175	145	235	165	195	115
5-E1	185	150	255	185	220	175
5-E2	200	180	245	170	215	120
5-E3	190	160	255	---	---	130

Aus der Lektüre des Wahrnehmungstests und der elektroglottographischen Aufnahmen haben sich folgende Zahlen ergeben. In der ersten Spalte geben wir Informant und Satz an; aus der zweiten kann man entnehmen, ob die Äußerung nach dem Wahrnehmungstest als ein Aussage- (A) oder als ein Fragesatz (F) bewertet wurde und der erreichte Prozentsatz; die dritte Spalte zeigt, wie hoch der Prozentsatz liegt, der die Äußerung als emphatisch erklärt. Schließlich sind die Hz-Schwankungen zu lesen:

a) Schwankung zwischen Anfangspunkt und erstem Tiefpunkt.

b) Schwankung zwischen erstem Tiefpunkt und erstem Gipfelpunkt.

c) Schwankung zwischen erstem Gipfelpunkt und zweitem Tiefpunkt oder Beugung, falls sie vorhanden sind.

d) Schwankung zwischen zweitem Tiefpunkt oder Beugung und Endpunkt, oder zwischen erstem Gipfelpunkt und Endpunkt, falls weder zweiter Tiefpunkt noch Beugung vorliegen.

Gleichzeitig haben wir die Sätze in drei Gruppen verteilt, je

nach den Ergebnissen des Wahrnehmungstests. Zuerst kommen die nicht emphatische, dann die emphatische Aussagesätze und schließlich die Fragesätze. Wir haben darin die fünf Informanten auseinander gehalten, um mögliche Idiolektalunterschiede berücksichtigen zu können.

Info-Satz	Aussage/ Fragesatz	Emphase	Hz-Schwankungen a	b	c	d
1-A1	A-100	33.3	-15	90	-70	-60
1-A3	A-100	33.3	-25	85	-95	-35
1-D1	A-100	5.6	-35	100	---	-115
1-D2	A-100	0	-20	55	-50	-65
1-D3	A-100	11.1	-50	85	-50	-75
1-E1	A-100	0	-15	40	---	-115
1-E2	A-100	5.6	-20	60	-55	-75
1-A2	A-100	55.6	-30	90	-65	-60
1-B1	A-94.4	76.5	-30	80	-85	-45
1-B2	A-100	72.2	-35	85	-85	-50
1-B3	A-100	88.9	-15	55	-90	-15
1-E3	A-100	66.7	-15	65	-95	-45
1-C1	F-88.9	25	-15	100	-95	90
1-C2	F-100	55.6	-30	105	-105	115
1-C3	F-94.4	35.3	-30	110	-130	95
2-A2	A-100	16.7	-35	140	-105	-60
2-B2	A-94.4	0	-60	125	---	-120
2-D1	A-100	5.6	-40	100	---	-125
2-D2	A-100	11.1	-40	95	---	-125
2-D3	A-100	16.7	-30	75	---	-120
2-E1	A-94.4	23.5	-35	100	---	-115
2-E2	A-100	5.6	-20	90	---	-145
2-E3	A-100	5.6	-20	80	---	-110
2-A1	F-94.4	52.9	-30	115	-110	95
2-A3	F-88.9	68.7	-20	115	-115	65
2-B1	F-94.4	64.8	-15	110	-105	110
2-B3	F-88.9	56.2	-20	95	-110	85
2-C1	F-88.9	68.7	-20	110	-110	115
2-C2	F-94.4	52.9	-40	115	-105	80
2-C3	F-100	83.3	-25	125	-120	95
3-A1	A-100	11.1	-20	95	-90	-50
3-A2	A-100	5.6	-30	85	-60	-35
3-A3	A-100	5.6	-10	75	-60	-55
3-D1	A-100	0	-20	70	---	-120
3-D2	A-100	0	-40	80	---	-125
3-D3	A-100	0	-20	85	---	-100
3-E1	A-100	0	-40	95	---	-105
3-E2	A-100	0	-15	85	---	-115
3-E3	A-100	0	-25	80	---	-115
3-B1	A-94.4	52.9	-10	65	-105	-10
3-B2	A-94.4	58.8	-25	60	-75	-20
3-B3	A-100	77.8	-15	55	-75	-35
3-C1	F-100	72.2	-10	85	-125	130
3-C2	F-94.4	70.6	-20	95	-95	100
3-C3	F-100	66.7	-5	85	-110	90
4-A1	A-100	27.8	-35	75	-65	-5
4-A2	A-100	11.1	-10	75	-50	-90
4-A3	A-100	16.7	-10	55	---	-40
4-B3	A-94.4	23.5	-25	80	---	-65
4-D1	A-100	0	-20	70	-30	-65
4-D2	A-100	5.6	-15	40	-40	-35
4-D3	A-100	5.6	-20	80	-50	-15
4-E1	A-100	5.6	-25	70	-40	-40
4-E2	A-100	5.6	-25	60	-45	-5
4-E3	A-100	0	-40	75	-50	-60

Info-Satz	Aussage/ Fragesatz	Emphase	Hz-Schwankungen a	b	c	d
4-B1	A-94.4	82.4	-25	65	-65	-35
4-B2	A-94.4	52.9	-20	60	-70	-35
4-C1	F-100	83.3	-35	45	-90	125
4-C2	F-88.9	93.7	-35	65	-100	140
4-C3	F-100	66.7	-35	40	-55	35
5-A1	A-100	27.8	-25	100	-85	-40
5-A2	A-100	38.9	-45	150	-105	-55
5-A3	A-100	11.1	-10	105	-105	-40
5-B3	A-94.4	29.4	-20	65	-60	-35
5-D1	A-94.4	5.9	-50	115	---	-115
5-D2	A-100	0	-10	45	-55	70
5-D3	A-100	5.6	-30	90	-70	-50
5-E1	A-100	0	-35	105	-70	-10
5-E2	A-100	5.6	-20	65	-75	-50
5-E3	A-100	27.8	-30	95	---	-125
5-B1	A-94.4	52.9	-40	105	-125	-15
5-B2	A-100	83.3	-35	100	-95	-25
5-C1	F-100	61.1	-25	85	-95	95
5-C2	F-100	72.2	-30	120	-120	145
5-C3	F-100	66.7	-40	120	-120	120

BESPRECHUNG

Es ist bemerkenswert, daß alle Sätze, die als Fragesätze bewertet, ebenfalls als emphatisch empfunden wurden, bis auf der 1-C1-Satz, dessen emphatischer Prozentsatz bei 25 liegt. Dieses einzige nicht emphatische Beispiel erlaubt uns jedoch nicht, irgendeinen Schluß zu ziehen, denn seine absoluten Hz-Werte genauso wie seine Hz-Schwankungen weisen keinen Unterschied auf in Vergleich mit den übrigen emphatischen Fragesätzen.

Dagegen ist der Unterschied zwischen Frage- und Aussagesatz eindeutig festzustellen, wenn man die d-Hz-Schwankungen beobachtet. In den Fragesätzen sind diese Schwankungen immer positiv und zwar ab 35 Hz im 4-C3-Satz, während sie in den Aussagesätzen immer negativ sind (ab -5 Hz in den 4-A1- und 4-E2-Sätzen). Es ist nur eine Ausnahme zu verzeichnen: in dem 5-D2-Aussagesatz ist diese Schwankung um 70 Hz positiv; wenn man diesen Satz mit den Fragesätzen vergleicht, kommt es vor, daß seine c-Hz-Schwankung wesentlich niedriger und nur vergleichbar mit der des 4-C3-Fragesatzes ist. Der Unterschied liegt hier wahrscheinlich in der Akzentstellung, denn, während der letzte Tiefton im 4-C3-Fragesätz mit dem letzten Wortakzent zusammenfällt, fällt dieser im 5-D2-Aussagesatz schon mit dem letzten Hochton zusammen, was ein weiterweisendes Signal darstellen mag.

Bei den emphatischen Aussagesätzen ist immer eine c-Hz-Schwankung zu beobachten, was nicht immer bei den nicht emphatischen Aussagesätzen der Fall ist. Man könnte also sagen, daß der Satz keine Emphase enthält, wenn die Tonhöhe vom Gipfelpunkt ausgehend allmählich bis zum Endpunkt absteigt.

Die emphatischen und die übrigen nicht emphatischen Aussagesätze unterscheiden sich dadurch, daß diese c-Hz-Schwankung in den emphatischen wesentlich größer ist als in den nicht emphatischen Sätzen: wenn wir zunächst mal die A-Sätze von der Untersuchung ausklammern, sind die c-Schwankungen in den emphatischen Sätzen im Vergleich mit denjenigen der nicht emphatischen Sätzen die folgenden:

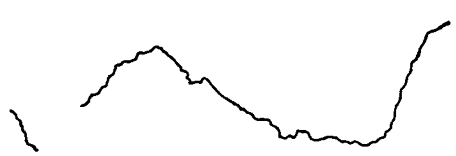
	emphatisch	nicht emphatisch	Unterschied
1. Informant:	ab -85 Hz	bis -55 Hz	30 Hz
2. Informant:	---	---	---
3. Informant:	ab -75 Hz	---	---
4. Informant:	ab -65 Hz	bis -50 Hz	15 Hz
5. Informant:	ab -95 Hz	bis -75 Hz	20 Hz

Es ist zu bemerken, daß die nicht emphatischen A-Sätze mehrmals einen nicht ganz eindeutigen Prozentsatz nach dem Wahrnehmungstest erweisen: 38,9% (5-A2), 33,3% (1-A1, 1-A3), 27,8% (4-A1, 5-A1), 16,7% (2-A2). Die Hz-Werte ihrer c-Schwankungen sind hier mehrmals auch ähnlich wie diejenigen der emphatischen Sätzen. Man kann auch dasselbe sagen vom einzigen emphatischen A-Satz (1-A2), der auch keinen sehr eindeutigen Prozentsatz erweist (55,6%) und einen niedrigeren Hz-Fall (-65 Hz) als die anderen emphatischen Sätze hat. Schließlich sind noch einige nicht emphatische A-Sätze (3-A1, 3-A2, 3-A3, 5-A3), die einen eindeutigeren Prozentsatz erweisen (11,1%, 5,6%) und auch einen starken Hz-Fall haben, und andere (4-A2, 4-A3), in denen keinen Unterschied im Vergleich mit den übrigen nicht emphatischen Sätzen zu erkennen ist. Der Unterschied zwischen den nicht emphatischen A-Sätzen und den übrigen emphatischen Sätzen scheint darin zu liegen, daß die Tonhöhe nach dem ersten Gipfelpunkt aufrechterhalten wird, um dann wieder von neuem die Intonation zu beginnen, nach einer mehr oder weniger kurzen Pause.

SCHLUSS

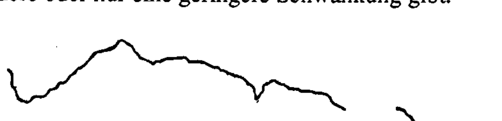
Nachdem wir gesehen haben, wo die Ähnlichkeiten und die Unterschiede liegen, können aus unserer Untersuchung folgende Schlüsse gezogen werden:

1. Die Fragesätze unterscheiden sich von den Aussagesätzen dadurch, daß die Hz-Schwankung zwischen dem zweiten Tiefpunkt oder der Beugung und dem Endpunkt aufsteigt; in den Aussagesätzen steigt sie immer ab, ob der Tiefpunkt oder die Beugung vorhanden sind oder nicht.

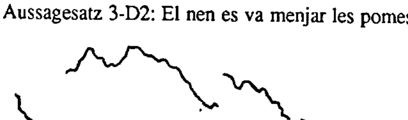


Fragesatz 5-C2: Les pomes es va menjar el nen?

2. Die nicht emphatischen Aussagesätze charakterisieren sich dadurch, daß die Tonhöhe nach dem ersten Gipfelpunkt allmählich bis zum Endpunkt absteigt, ohne daß es sich keine besondere oder nur eine geringere Schwankung gibt.

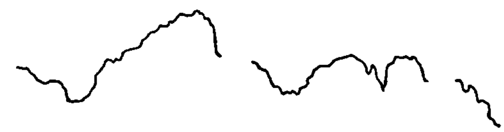


Aussagesatz 3-D2: El nen es va menjar les pomes.



Aussagesatz 5-D3: El fred va matar els arbres.

3. Einige nicht emphatische Aussagesätze, die ja syntaktisch genau charakterisiert werden können, erweisen trotzdem eine größere Schwankung nach dem ersten Gipfelpunkt. Diese Schwankung findet aber normalerweise nach einer kurzen Pause statt, und erfolgt nur bei thematisierten Sätzen.



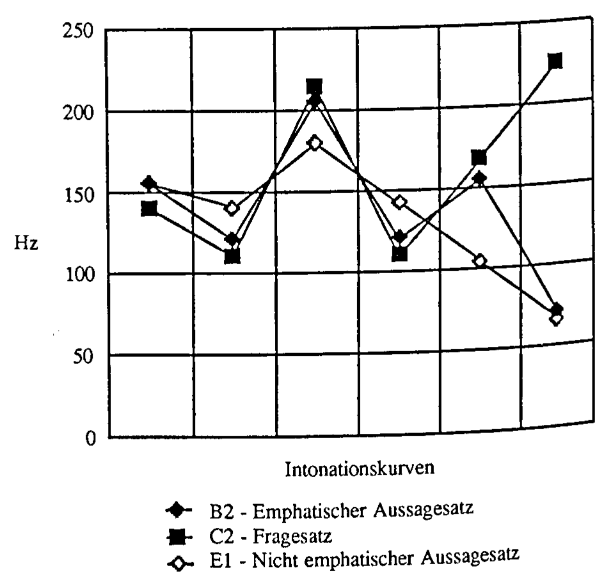
Aussagesatz 5-A3: Els arbres, els va matar el fred.

4. Die emphatischen Aussagesätze charakterisieren sich dadurch, daß die Tonhöhe einer großen Hz-Schwankung nach dem ersten Gipfelpunkt unterliegt, ohne daß eine Pause dazwischen liegt.

Aussagesatz 3-B3: Els arbres va matar el fred.

Schließlich ist es auch hinzuzufügen, daß diese Beobachtungen wahrscheinlich nicht von den anderen suprasegmentalen Elementen (Tempo und Intensität) getrennt werden können. Während der ganzen Untersuchung haben wir die Vermutung gehabt, daß Intensität und Tonhöhe sich ergänzen: wo die Tonhöhe in einer emphatischen Satz nicht so charakteristisch wirkt, ist die Intensität stärker, oder umgekehrt. Tempo und Tonhöhe mögen sich auch wohl ergänzen: den Tiefpunkt oder die Beugung vor dem Satzende wird um so schneller erreicht, desto größere die Hz-Schwankung ist. Diese Verhältnisse müssen aber noch weiter untersucht werden.

Vergleichung am Beispiel einiger Sätze des 1. Informanten



PERCEPTUAL ASPECTS OF EMOTIONAL SPEECH

MALL LAUR

Information and Computer Centre
State Committee of TV and Broadcasting Centre of the ESSR
Tallinn, Estonia, USSR 200100

ABSTRACT

The verbal aspect of short utterances can affect the perceptual process of emotional speech. The affect is observed on the basis of three different stimuli, a short greeting among them. The investigation is cross-cultural.

INTRODUCTION

The perception of emotions by vocal cues has been examined by several authors. Among the factors which might affect the perception appear to be sex of listener /8/, age of listener /4/,/5/,/7/, cultural distance between speaker and listener/10/, /1/,/2/,/3/. The verbal aspect of speech stimuli has often been eliminated /6/.However, aiming to investigate emotion perception in the process of speech communication, the verbal aspect of stimuli may not be neglected.

The goal of the present research was to examine emotion perception on the basis of short utterances (mono- and disyllabic sentences, 260-360 ms in neutral speech). Short duration of a signal may cause deficiency of vocal cues and subsequently listener's perception can be affected by the verbal aspect.

METHOD

Stimuli

Three stimuli in Estonian were selected. To check the insufficiency of a short utterance for emotion perception, a four word utterance was chosen for one stimulus (1)"Taavi saatis Saarale kaardikese." - "David sent a card to Sarah", later referred as 'long sentence' or LS.

Short utterances differed in their meaning and position in a dialogue:

(2)"Tere" - "Hello", a most common greeting in Estonian; later referred as 'greeting' or G.

(3)"Saab" - 3rd p. sing. pres. indicative of the verb 'saama' meaning 'to get,obtain or receive sth; to become sth, sb; used both as a personal and an impersonal predicate; later as 'short sentence' or SS.

Emotions from Izard's study /9/ - surprise, interest, joy, fear, sadness, shame anger, contempt, disgust - and in addition love and neutral were chosen for emotional categories.(Disgust was not used for greeting).

Recordings were made in a soundproof booth using a microphone connected to a tape recorder outside the booth.

Subjects and the Procedure

The stimuli, set in a random order, were rendered twice. During the first session listeners had to label the emotions. At the second session they had to choose a response out of the 10 or 11 categories. The first test will be referred as 'free choice test', the second as 'forced choice test'.Pauses for responding lasted ten seconds, the sequence number of a stimulus was checked after every 5 stimuli. When the primary group had accomplished both tests, the stimuli were presented in a rearranged order to a control group.

A part of the stimuli (28 long sentences, 27 short sentences and 40 greetings) were rendered 28 Russians from Moscow State University (students and the staff, no knowledge of Estonian) to accomplish the forced choice test.

The sizes of Estonian listener groups: for long and short sentence 65 subjects in the primary group and 21 subjects in the control group; for greeting 48 subjects in the primary and 28 subjects in the control group; the division between genders was roughly half in all subject groups.

RESULTS AND DISCUSSION

The results of forced choice test form the basis of the following discussion. Overall mean of identification scores of Estonian subjects for the three groups of stimuli, long sentence, short sentence and greeting, did not differ ($\bar{x}$=49,2;49,4; 42,7 accordingly, see Table 1).Still, the comparison by categories revealed some differences (see Table 2).

TABLE 1. Mean percentage of correct identifications of emotional categories by Estonian subjects.

	LONG SENT. N=65+21		SHORT SENT. N=65+21		GREETING N=48+28	
	1	2	1	2	1	2
neutr.	6	88.7	4	75.0	4	61.1
surprise	6	58.4	4	40.4	4	58.9
interest	3	27.1	4	55.5	4	34.2
joy	6	61.0	4	61.2	4	45.6
love	6	63.4	4	62.6	4	53.5
sadness	6	59.4	4	47.8	4	47.4
fear	6	52.3	2	44.7	4	30.9
shame	3	23.1	3	14.6	4	20.7
anger	6	40.8	4	59.6	4	36.9
contempt	6	50.0	2	56.3	4	38.9
disgust	3	17.1	3	25.7		
overall	57	49.2	35	49.4	40	42.7

1 - number of stimuli
2 - % of correct identifications

TABLE 2. Analysis of variance by means of the T-test between the identification scores of emotional categories of different stimuli.

	LS × SS		LS × G		SS × G	
	T	df	T	df	T	df
neutr.	3.602	11[x]	4.321	8[x]	1.901	12
surprise	1.893	15	0.013	15	1.666	16
interest	5.901	12[x]	0.893	11	2.758	10[x]
joy	0.033	16	1.326	11	1.260	13
love	0.997	18	0.875	18	0.763	16
sadness	1.498	19	1.352	16	0.073	15
fear	0.669	5	3.557	20[x]	1.190	5
shame	1.610	15	0.847	5	1.330	14
anger	2.903	20[x]	0.491	18	3.229	14[x]
contempt	0.809	11	1.481	19	2.219	10[x]
disgust	0.207	7				
overall	0.041	43	1.191	37	0.996	20

[x] $p<0.05$

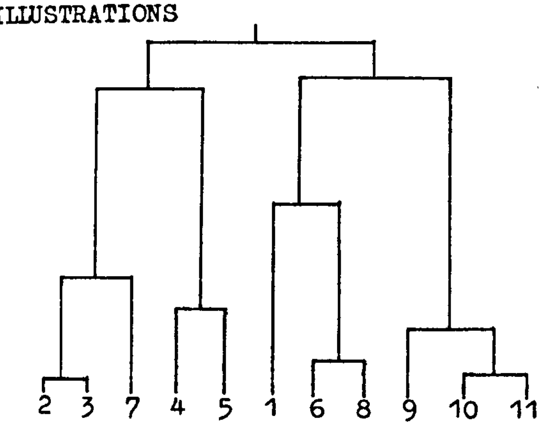
TABLE 3. Mean percentage of correct identifications of emotional expressions by Russian and Estonian subjects.[x]

	LONG SENT.		SHORT SENT.		GREETING	
	Russ. N=28	Eston. N=21	Russ. N=28	Eston. N=21	Russ. N=28	Eston N=48
neutr.	83.3	85.7	70.2	76.0	49.1	61.5
surprise	35.3	68.3	54.7	58.7	41.1	66.2
interest	17.9	61.9	10.7	63.6	20.6	35.8
joy	59.5	69.8	71.4	61.9	20.5	41.2
love	63.6	77.8	80.9	82.2	44.6	51.6
sadness	71.4	68.3	44.6	56.0	50.9	41.2
fear	57.1	52.4	62.5	54.7	20.6	31.3
shame	33.6	21.4	21.4	21.6	8.1	20.9
anger	66.2	41.3	48.8	61.9	48.2	38.9
contempt	66.2	65.6	65.5	61.9	49.1	36.5
overall	55.4	61.3	52.7	59.9	35.3	42.5

[x]Responses of these Estonian groups have been considered who accomplished the test in equal conditions (sequence of stimuli was the same): control group for LS and SS, primary group for G.
The stimuli in Table 3 do not entirely coincide with those reported in Table 1.

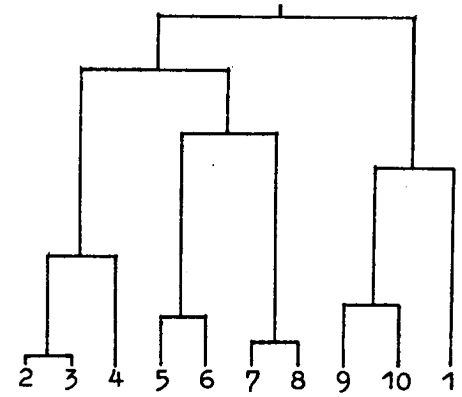
The comparison of identification scores of Estonian listeners did not reveal the suppositional affect of verbal aspect of short utterances on emotion perception - the overall mean scores were similar and the differences on category level did not yield any regularity.
Cluster analysis /11/, carried out on the confusion matrices demonstrated that on the basis of long sentence, emotion perception had proceeded from the conceptual, positive - negative dimension. That holds true for both groups of listeners, Estonians and Russians, i.e. the verbal aspect of a longer utterance did not have any affect on emotion perception (see Fig. 1 and Fig.4).
The confusions occurred in the responses of Estonian listeners to short utterances, SS and G, revealed a discrepancy - the regular confusion of surprise with interest was missing in the responses to greeting, interest had been included into the cluster of passive emotions (see Fig.2 and 3); surprise had been confused with joy. The confusion clusters of G can easily be explained if the verbal aspect of this stimulus is taken into consideration. A listener, hearing a greeting, is foremost interested in the probability of conversation continuation. If the speaker seems to be pleasantly surprised (surprise+joy), conversation will most likely follow. If the greeting is purely formal (neutral, contemptuous, angry), no conversation is expected. If the greeting expresses speaker's passiveness (passive emotions), the continuation of conversation will depend on the listener.
The described affect of verbal aspect can be confirmed if the responses of Russian listeners reflected a different attitude, in fact they did. The responses of Russian listeners to short sentence and greeting revealed rather unity than discrepancy (see Fig.5 and 6) - in both samples active positive and negative emotions had been confused; the confusion of surprise and interest is present in both dendrograms. The comparison of the identification percentages of Russian and Estonian listeners at the category level yielded another evidence in favour of the affect of verbal aspect on emotion perception on the basis of short utterances. Namely, quite unexpectedly an association between a meaning of the stimulus (short sentence) and an emotional category (interest) had occurred - "saab" could be interpreted as "Interest in whether sth can be obtained". As a re-

ILLUSTRATIONS



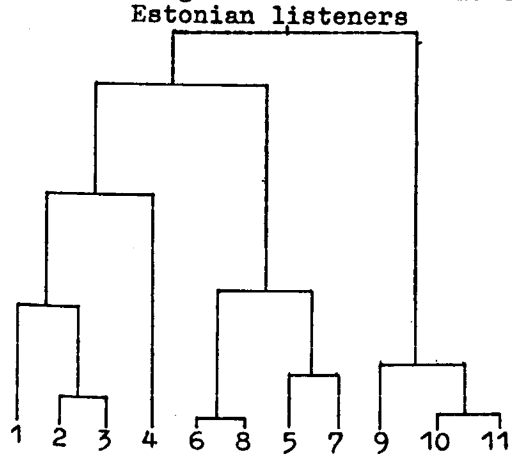
1 neutral 4 joy 7 fear 10 contempt
2 surprise 5 love 8 shame 11 disgust
3 interest 6 sadness 9 anger

Fig. 1. Long sentence. Confusions of Estonian listeners



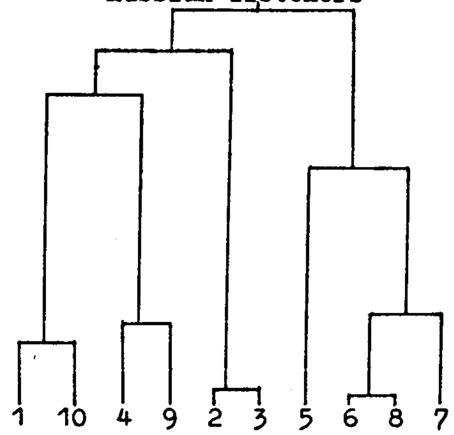
1 neutral 4 joy 7 fear 10 contempt
2 surprise 5 love 8 shame
3 interest 6 sadness 9 anger

Fig. 4. Long sentence. Confusions of Russian listeners



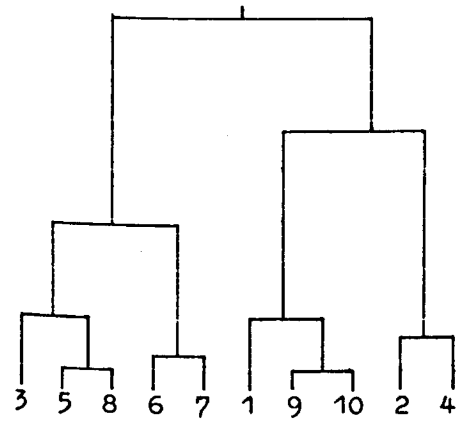
1 neutral 4 joy 7 fear 10 contempt
2 surprise 5 love 8 shame 11 disgust
3 interest 6 sadness 9 anger

Fig. 2. Short sentence. Confusions of Estonian listeners



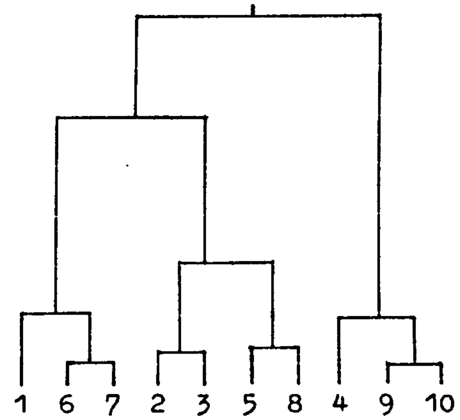
1 neutral 4 joy 7 fear 10 contempt
2 surprise 5 love 8 shame
3 interest 6 sadness 9 anger

Fig. 5. Short sentence. Confusions of Russian listeners



1 neutral 4 joy 7 fear 10 contempt
2 surprise 5 love 8 shame
3 interest 6 sadness 9 anger

Fig. 3. Greeting. Confusions of Estonian listeners



1 neutral 4 joy 7 fear 10 contempt
2 surprise 5 love 8 shame
3 interest 6 sadness 9 anger

Fig. 6. Greeting. Confusions of Russian listeners

sult, interest had been well identified on the basis of short sentence by Estonian listeners whereas Russian listeners had not distinguished interest at all rating all these stimuli to express surprise.
The evidence supporting the hypothesis about the affect of verbal aspect on emotion perception is not strong - the emotional category of interest the perception of which forms the basis for this evidence is too ambiguous and the present argument may turn to be wrong. Thus further research in this direction - investigation of verbal aspect on emotion perception in different conditions, different speech signals - is necessary either to confirm the hypothesis under discussion or to disprove it.

CONCLUSIONS

Long utterances used in this research as stimuli (a sentence of four words: 2 disyllabic, 1 trisyllabic and 1 four-syllabic word) favoured emotion perception. The effect became evident in both groups of listeners - who understood the stimuli (Estonians) and who did not understand (Russians).

The presumable affect of verbal aspect of short utterances (mono- and disyllabic sentences) on emotion perception became manifest mostly through different perception of interest.

REFERENCES

/1/ Albas, D.C., McCluskey, K.W., Albas, C.A. "Perception of the emotional content of speech: a comparison of two Canadian groups."- Journal of Cross-cultural Psychology, 1976, 7, 481-490.

/2/ Beier, E.G., Zautra, A.J., "Identification of vocal communication of emotion across cultures." - Journal of Consulting and Clinical Psychology, 1972, 39, 166.

/3/ Bezooijen, R. van. "Recognition of Dutch expressions of emotions in Taiwan."- Proceedings of the Institute of Phonetics, Nijmegen, 1982, 6, 1-8.

/4/ Bezooijen, R. van. "Recognition of vocal expressions of emotion by toddlers." - Proceedings of the Institute of Phonetics, Nijmegen, 1983, 1, -8.

/5/ Davitz, J.R. "Auditory correlates of vocal expressions of emotional meaning." - In Davitz, J.R. (Ed.) "The communication of emotional meaning." Wesport (Ct.): Greenwood Press, 1964, 101-113.

/6/ Davitz, J.R., Davitz, L.J. "The communication of feelings by content-free speech."-Journal of Communication, 1959, 9, 6-13.

/7/ Dimitrovsky, L. "The ability to identify the emotional meaning of vocal expressions at successive age levels." - In Davitz, J.R. (Ed.) "The communication of emotional meaning." Westport (Ct.): Greenwood Press, 1964, 69-87.

/8/ Hall, J. "Gender effects in decoding nonverbal cues." -Psychol. Bulletin, 1978, 85, 845-857.

/9/ Izard, C.E. "The face of emotion." New York; Appleton Century Crofts, 1971.

/10/ Kramer, E. "Elimination of verbal cues in judgments of emotion from voice."- Journal of Abnormal and Social Psychology, 1964, 68, 390-396.

/11/ Orloci, L. "An agglomerative method for classification of plant communities." Journal of Ecology, 1967, 55, 193-205.

COMPUTER ASSISTED DIAGNOSIS OF PERCEPTUAL ERRORS

RUDOLF WEISS

Department of Foreign Languages
Western Washington University
Bellingham, Washington USA 98225

ABSTRACT

A computer program has been developed for the scoring and analysis of perceptual errors in classifying German vowels. The program, written in "BASIC" for MS-DOS system computers, plots out specific errors and provides an accuracy index and length agreement correlate. A second part of the program provides the learner with a ranking list of specific vowel difficulty and an explanation of the likely nature of the perceptual error. The results may either be printed or viewed on the screen.

INTRODUCTION

The author has for some time been concerned with studies of perception, in particular its application to corrective procedures with the ultimate goal of correcting and improving pronunciation of learners of German. It has long been the author's belief that errors of pronunciation and errors of perception go hand in hand and that correction of both perception and production must be addressed. This has been the subject of several earlier papers ([1, 2]) and is the underlying premise of a book co-athored by H.-H. Wängler which has recently been published by Western Washington University Press [3] and is now used as a text by a number of German departments in the USA.

The contrastive phonetic approach used in the book is ideally suited for computer application. Each sound is treated individually with a number of pedagogically oriented steps provided to facilitate mastery of the sound in context based on potential perception and articulation difficulties. A perceptual or listening frame with accompanying listening tests in each case precedes actual production exercises. The listening exercises set a framework for contrastive problems both between potentially conflicting native (L_1 = English) as well as target (L_2 = German) sounds and contexts. The predetermining factors as the potential of likely problems for each sound are based upon contrastive phonetic principles and upon data gathered in the past administration of a perception test developed for native German speakers and then modified for non-native learners [4].

The test which has been modified numerous times has served in the past as an accurate indicator of degree of nativeness in perception. It is comprised of minimal pairs containing variations of German vowels which are then classified as one of fifteen phonemic categories in German. The test has in the past yielded valuable data about ranking order of vowel difficulty for students at various levels of study and has provided numerical indexes corresponding to performance standards for levels from first year college to advanced graduate student status [5].

However in its specific application here, the test is seen as an invaluable aid as part of a basic program aimed at improving individual language skills. This is done by administration of the test at varied intervals noting specific progress at elimination or improvement of certain perceptual errors. The computer program is designed to indicate specific perceptual errors, provide a priority listing of most frequently made errors and the likely nature of both errors affecting the general classification (or misperception) of vowel categories as well as specific vowel errors. As such the program has proved to be a valuable learning tool facilitating more automatic and accurate assessment of difficulties and has applications which greatly facilitate computer-dependent learner acquisition of sound perception/production.

EXPERIMENTAL PROCEDURES

The test was administered individually via a Tandberg Model 812 cassette recorder and headphones linked to an IBM-PC by a serial connection. The test material is displayed for the subject on a Teknika MJ-22 RGB Monitor or may be printed on an Epson LQ-1500 or FX-80 printer. The equipment is housed in the Foreign Language Learning Center at Western Washington University.

The student must classify each of 100 items on tape as one of fifteen phonemic choices. These choices appear as orthographic representations. The choices are indicated as letters A through O. At the conclusion of the test the student is provided with a display of all errors made along with a general assessment of major perceptual errors (6). The student may review the errors on the screen or receive a printed hard copy via printer as shown in Figures 1 and 2.

Figure 1

NAME: 1286elm31
Level 314
DATE: 12/86

ITEM	TOTAL	INCORRECT RESPONSES NO.	SUBSTITUTIONS MADE	LAF YES	LAF NO
[i:]	6	1	[y:]- 100	1	0
[I]	3	1	[e:]- 100	0	1
[e:]	7	4	[i:]- 50 [ɛ:]- 25 [ɔ]- 25	3	1
[ɛ]	1	0		0	0
[ɛ:]	11	9	[e:]- 22 [a:]- 22 [a]- 33 [ø:]- 11 [ɛ]- 11	5	4
[a:]	15	3	[ɔ]- 66 [a]- 33	0	3
[a]	2	1	[ɔ]- 100	1	0
[u:]	6	1	[o:]- 100	1	0
[U]	4	2	[œ]- 50 [ɔ]- 50	2	0
[o:]	12	0		0	0
[ɔ]	3	0		0	0
[y:]	10	7	[ø:]- 28 [Y]- 28 [œ]- 14 [e:]- 14 [ɛ:]- 14	4	3
[Y]	0	0		0	0
[ø:]	16	5	[œ]- 40 [y:]- 40 [u:]- 20	3	2
[œ]	5	4	[U]- 100	4	0
TOT.	100	38	TOTAL PERCENTAGE OF CORRECT RESPONSE: 62	24	14
			TOTAL PERCENTAGE OF LAF: 63		

Figure 2

NAME: 1286elm31
VOWELS IN ORDER OF DIFFICULTY:

VOWEL		ERRORS
[ɛ:]	-	9 OF 11 OR 81 %
[œ]	-	4 OF 5 OR 80 %
[y:]	-	7 OF 10 OR 70 %
[e:]	-	4 OF 7 OR 57 %
[a]	-	1 OF 2 OR 50 %
[U]	-	2 OF 4 OR 50 %
[I]	-	1 OF 3 OR 33 %
[ø:]	-	5 OF 16 OR 31 %
[a:]	-	3 OF 15 OR 20 %
[i:]	-	1 OF 6 OR 16 %
[u:]	-	1 OF 6 OR 16 %

GENERAL OBSERVATIONS:

LAF = 63 % - TOO MUCH DEPENDENCE ON LENGTH FACTOR
NOT ATTENTIVE ENOUGH TO QUALITY DISTINCTIONS AMONG VOWELS

6 ERRORS OR 15% OF THE ERRORS ARE DUE TO AN INABILITY TO CLEARLY DISTINGUISH BETWEEN UMLAUTED AND NON-UMLAUTED VOWELS. BE ATTENTIVE OF ORTHOGRAPHIC DIFFERENCES BETWEEN A/Ä, U/Ü, ETC.

2 ERRORS OR 5% ARE DUE TO THE INABILITY TO CLEARLY DISTINGUISH BETWEEN LIP-ROUNDED AND NON LIP-ROUNDED FRONT VOWELS. BE ATTENTIVE OF THE DIFFERENCE BETWEEN [i:]/[y:], [e:]/[ø:], ETC.

9 ERRORS OR 23% OF THE ERRORS WERE DUE TO THE SUBSTITUTION OF ONE UMLAUTED VOWEL FOR ANOTHER. THIS INDICATES THAT THERE IS SOME CONFUSION IN DISTINGUISHING ONE UMLAUTED VOWEL FROM ANOTHER. BE ATTENTIVE TO THE DIFFERENCE BETWEEN Ö/Ü VOWELS.

COMPUTER PROGRAM

The computer program written for this application is in two parts. The first part generates on-screen directions for taking the test and generates data files through a sub-routine of test responses. The responses themselves are converted from letters A through O accessed on the key-board to numerical values 1 through 15.

The second part of the program is the analysis routine [7]. It is written in BASIC for MS-DOS with sub-routines compiled in machine language to increase response time. It comprises two major sections. The results of the first section are illustrated as Figure 1. The program first performs a matching function comparing the data file generated by the student with the data file of the key of correct responses. Sub-routines perform the statistical functions of calculating the errors made. The initial analysis compiles an error index for each vowel based upon the agreement factor with the individual vowel. A total percentage for the test is calculated. A second sub-routine in the program classifies each vowel as a subset of either a short vowel group or long vowel group and calculates errors on the basis of whether they are in agreement with the length or in disagreement. The extent of this agreement is calculated as the LAF (length agreement factor). Further sub-routines classify the errors and create a hierarchical arrangement of the errors for individual vowels along with the percentage of the frequency of that error for the specific vowel.

The display of errors as indicated in Figure 1 are in phonetic symbols and may be displayed either on the screen or printed. The screen program is accomplished through a screen sprite routine using an IBM character generator. The printer routine utilizes graphics characters generated through Printworks [8] graphics program and downloaded to the internal buffer of the printer.

The basic display of errors and statistical analysis is followed by a second section which provides more directed diagnostic help to the learner based on further analysis of the errors. The results of the second phase of analysis are indicated as Figure 2. The types of errors are reclassified to provide more specific diagnostic help aimed at assisting the student to improve his/her perception. First a listing of vowels is provided, arranged in terms of perceptual difficulty for the student. The number of errors compared to the total number of that specific vowel contained on the test is indicated along with a percentage of misclassification of that vowel.

This is followed by a section called "General Observations" and is again comprised of a number of sub-routines comparing errors to specific arrays of character strings. The first statement provides an analysis of the LAF mentioned previously. Since the test items were intended to exhibit deliberate manipulation of both the quality and length axis, the errors should have been roughly divided evenly between length and quality, an LAF of 50% ± 10% would thus be considered within the norm. If the LAF is less than 40%, the LAF percent factor is indicated along with the statement "Wrong length substituted—Not attentive enough to length differentiation among vowels." If on the other hand the LAF is greater than 60%, a statement such as that in Figure 2 appears indicating that too much dependence was placed upon length in classifying vowels and not enough upon qualitative distinctions. Further routines in this part of the program compare errors as character strings to distinguish between umlauted vs. non-umlauted sounds (indicating possible orthographic interference), lip-rounded vs. non lip-rounded and umlauted vs. other umlauted vowels. These categories usually account for approximately 50% of all student perception errors.

The following would serve as an example of the nature of a small segment of the analysis routine. A statement intending to express the substitution factor of lip-rounded for non lip-rounded vowels and vice versa would use as a basis the mutual substitutions of y:/Y/ø:/oe for i:/I/e:/ɛ and vice versa. The letter codes would analyze substitutions of ABCD for LMNO and vice versa, whereas the routine would identify them as numbers 1,2,3,4 for 12, 13, 14, 15 and vice versa. The complete statement for this routine is given below as lines 6540 through 6570 as it actually occurs in the program.

List: 6540-6570

```
6540 IF ER>0 then if O$="p" OR O$="P" THEN LPRINT
AI$:LPRINT A2$:LPRINT A3$:LPRINT ELSE PRINT
AI$:PRINT A2$:PRINT A3$:PRINT

6550 ER=CRI(I,12) + CRI(I,13) + CRI(I,14) + CRI(I,15)
+ CRI(2,12) + CRI(2,13) + CRI(2,14) + CRI(2,15) + CRI(3,12)
+ CRI(3,13) + CRI(3,14) + CRI(3,15) + CRI(4,12) + CRI(4,13)
+ CRI(4,14) + CRI(4,15)

6555 ER=ER+CRI(12,I) + CRI(12,2) + CRI(12,3) + CRI(12,4)
+ CRI(13,I) + CRI(13,2) + CRI(13,3) + CRI(13,4) + CRI(14,I)
+ CRI(14,2) + CRI(14,3) + CRI(14,4) + CRI(15,I) + CRI(15,2)
+ CRI(15,3) + CRI(15,4)

6560 AI$=STR$(ER)+" ERRORS OR" + STR$(INT
(ER/NW*I00)) +"% ARE DUE TO THE INABILITY
TO CLEARLY DISTINGUISH": A2$="BETWEEN LIP-
ROUNDED AND NON LIP-ROUNDED FRONT VOWEL
S. BE ATTENTIVE"

6570 A3$="OF THE DIFFERENCE BETWEEN [i:]/[y:],
[e:]/[" + CHR$ (SCR(I4)) + ":], ETC."
```

The program has been further developed to provide a more detailed diagnostic analysis of individual vowels. A student can choose to review the errors for individual vowels. The most common errors indicated as substitutions in Figure 1 are then diagnosed in detail along with the severity of that error. For example if [i:] were perceived as [I] a statement indicating that the long vowel (bieten was perceived as short (bitten) would appear; if [i:] were perceived as [e:] a statement would ensue indicating that the perception was one of the wrong quality (beten instead of bieten); or if [i:] were perceived as [ɛ] a statement would follow indicating that a long high vowel was perceived as a short vowel of lower quality (betten instead of bieten). In this fashion errors reflecting all commonly substituted vowels are given brief explanations as to the nature of the error.

RESULTS

This analysis program has provided a useful tool in attempts to correct perception errors. It affords the possibility of self-administration of the test and repeated attempts at frequent intervals to monitor progress towards the elimination of errors. It furthermore allows the opportunity to concentrate efforts in goal-directed fashion on specific perceptual problem areas. Since the nature of the errors are by and large predictable based upon contrastive phonetic distinctions between English and German, this program could be further enhanced by providing moving graphic illustrations on the screen correlative to specific physiological activity produced in yielding the error. The program also has the potential coupled to a digitizing/synthesizing package to serve as a precursor to provide virtually automated recognition and correction of pronunciation errors. Together with an interactive video display the result ultimately could be a "computerized phonetician," at least within a limited context where errors are relatively predictable.

REFERENCES

[1] Weiss, R. "The Role of Perception in Teaching the German Vowels to American Students," Proceedings of the IVth International Congress of Applied Linguistics, Vol. 3, Stuttgart: Hochschul Verlag, 1976, pp. 513-524.

[2] Weiss, R. and H.-H. Wängler, "Über das Unterrichten deutscher Vokalwerte auf der Grundlage perzeptorischer Normen," Forum Phoneticum, Vol. 5, Hamburg: Helmut Buske Verlag, 1978, pp. 63-78.

[3] Weiss, R. and H.-H. Wängler. German Pronunciation: A Phonetics Manual, Bellingham: Western Washington University Press, 1985, 386 pp.

[4] Weiss, R. "Perception as an Aid in Teaching Pronunciation," Proceedings of the IXth International Congress of Phonetic Sciences, Vol. I, Copenhagen: Institute of Phonetics, University of Copenhagen, 1979, p. 426.

[5] Weiss, R. "A Perception Test as a Diagnostic Tool in Teaching German Pronunciation," Current Issues in the Phonetic Sciences II, (ed. by H. and P. Hollien), Amsterdam: John Benjamin, B.V., 1979, pp. 905-916.

[6] A second version of the test is now almost complete which allows specific vowels to be isolated and provides immediate feedback of errors. It is expected that this version will be particularly useful for an experiment to determine whether goal-directed practice of perception has a significant effect upon changing perceptual parameters.

[7] The analysis routines and sub-routines were developed according to the author's specifications by Scott Honaker, a computer programming student at Western Washington University.

[8] Printworks is a registered trademark of SoftStyle, Inc., 7192 Kalanianaole Hwy., Suite 205, Honolulu, Hawaii 96825. For this application Version 1.0 (Copyright 1984) was used.

TEACHING PHONETICS USING THE PHONETIC DATA BASE ON MICROCOMPUTER

JOHN H. ESLING

Department of Linguistics
University of Victoria
Victoria, B.C. V8W 2Y2 Canada

ABSTRACT

Development of a microcomputer-based speech processing, analysis, manipulation and input/output system allows the incorporation of revised techniques in the teaching of basic phonetics. The Micro Speech Lab and related speech editing software permit acquisition, storage, random-access retrieval, variable-order selection, marking, concatenation, and auditory and visual comparison of phonetic data. Phonemic inventories and speech samples of several diverse languages illustrating a variety of phonetic contrasts have been collected for research and instructional applications in a Phonetic Data Base. Tasks are described which give phonetics students the opportunity to collect speech sound data, hear and evaluate linguistic and indexical contrasts, and extract short samples for illustration, comparison and practice.

THE MICRO SPEECH LAB SYSTEM

The procedures for phonetics instruction described here are the direct result of the development of a Micro Speech Lab microcomputer-based system for capture, playback and analysis of speech and other acoustic signals. Micro Speech Lab (MSL) is a complete hardware/software package for use with IBM (PC, XT, AT) microcomputers, designed and developed in the Centre for Speech Technology Research/ Phonetics Laboratory at the University of Victoria. MSL contains a software diskette and internally mounted data acquisition hardware including anti-aliasing filters, A/D and D/A circuitry, and a user's manual with user instructions and descriptions of theory of use and applications [1]. The software includes user control of signal input, several waveform displays, audio output, analysis (amplitude, pitch, spectrum) and file management. Phonetic instruction using MSL applies this rapid random-access speech input/output capability to the recording, storage, recall, comparison, visual waveform observation and manipulation, spectral analysis, and variable auditory presentation of speech sound material collected in a Phonetic Data Base.
In addition, a program written to supplement MSL's speech-capturing, storage and processing capabilities, MSL EDIT, allows students to access and display graphic waveforms of sampled data files in order to listen to words or several-second samples of text in any language selected, vary listening sequences, edit existing files, and combine elements of old files into new files. "Designed as a supplementary package to accompany the Micro Speech Lab, the purpose of the program is to provide a highly flexible method for auditory examination and manipulation of digitally stored signals" [2]. Up to five sampled data files can be displayed and monitored individually, in reverse, or in continuous repetition of sequences composed of parts of any file.

THE PHONETIC DATA BASE

To provide a core of linguistically organized speech data for instructional and research purposes, a Phonetic Data Base of speech samples has been assembled using MSL. Words and text drawn from numerous linguistic, sociolinguistic and dialect survey sources, represent a wide range of speech sounds of languages of the world. Samples are digitally encoded using the MSL capturing routine on the IBM-PC microcomputer. Files are stored by language on diskette or hard disk and documented on paper by number for reference to phonetic, phonemic and orthographic representations and English gloss of each sample.
Examples of phonetic sounds that are normally difficult to obtain, and phonemic inventories of a range of languages not usually encountered or available during the course of most phonetics classes have been included. Languages collected thus far include: Egyptian Arabic, Inuktitut, Korean, Miriam, Nitinaht, Nyangumarta, Rutooro, Scots Gaelic, Skagit (Coast Salish), Spokane, Turkish, Umpila, Xhosa, and Yoruba. At least 50 words and several short text files for each language have been stored in the current library.
This system has been made available to students in phonetics classes, including those in Applied Linguistics (Teaching English as a Second Language) teacher preparation programs. Individual words and short texts are accessed from diskette or hard disk directories and manipulated by groups of students according to tasks set by the instructor to focus on particular auditory categories. Most tasks are carried out with MSL EDIT, and most are performed outside of class time.

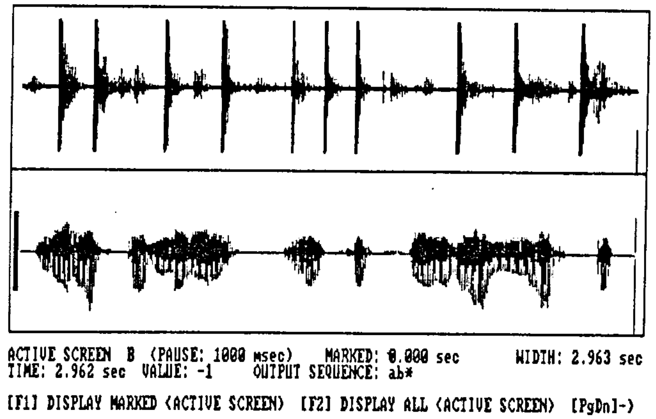
REVISING METHODS OF INSTRUCTION

The Phonetic Data Base (PDB) is intended to provide a practical, accessible and realistic mechanism for experiencing, comparing and evaluating the range of the multidimensional acoustic space reflected in the phonetic chart. The PDB gives manipulative and creative power to learners [3], and illustrates to prospective teachers how students can be enabled to collect and store language items in a format that allows easy reorganization [4] [5].
The goal is to enhance phonetics instruction by incorporating the PDB and MSL delivery system with a number of recent developments in second language acquisition theory into the structure of the phonetics course. Revised techniques emphasize the role of prosody, including voice setting, in the initial stages of phonetic exposure rather than focusing attention immediately on segmental analysis. Attention is given to the interpretation of indexical as well as of phonological properties of speech, for listeners first encountering a new language.
Auditory recognition and assignment of written symbols to represent categories of sounds are the central skills to be developed here, as they are in second language (L2) listening/speaking tasks that emphasize aural discrimination rather than production [6] [7]. Many current, popular L2 teaching approaches omit explicit teaching of pronunciation [8] [9] and may leave language teachers with no clear model of how to present L2 speech sounds other than their recollection of how they themselves were taught phonetics. This decrease in overt attention given to the pronunciation component of L2 teaching has caused some alarm [10], and it is hoped that this discrepancy can be partially reduced by introducing a modified approach into the course where language teachers originally learn phonetics.
The emphasis in L2 teaching is shifting away from the static model approach based on the ideal phonemic inventory of the target language taught in a dedicated pronunciation class, towards communicative, problem-solving task-based activities designed to provide larger amounts of L2 for manipulation by students [11]. Where pronunciation is taught explicitly in L2 programs, the focus has shifted to word-level meaning contrasts rather than phoneme drills, and to the early introduction of prosodic features [12] [13] [14] [15]. Specific conditions found to benefit L2 acquisition and teaching include: (1) diversity of language material presented in meaningful situations, (2) experience and practice in perception before production is required, (3) clear identification and association of concrete referents, assimilated at the student's own pace, (4) presence of significant target language models, especially of peers [16].
The benefit for phonetics teaching is the range of "dynamic variety" that can be presented with the random-access capability of MSL. Instead of relying on imitation and memorization in production of segmental units, MSL allows students to satisfy perceptual criteria by obtaining understandable "input" of sounds in their natural context before having to produce "output" [17]. Students can choose their own pace and sequence using MSL, rather than having recorded examples presented in uncontrolled order. The criterion of identification with speaker models is met as students use MSL to choose specific individuals' voices to listen to and arrange spoken material into phonetic classification schemes [18]. In this way, students gradually build an inventory of sounds and symbols, to complete the phonetic chart, based on "input" which they have collected and classified themselves according to the phonetic principles they are learning.

TASK-ORIENTED INSTRUCTION IN BASIC PHONETICS

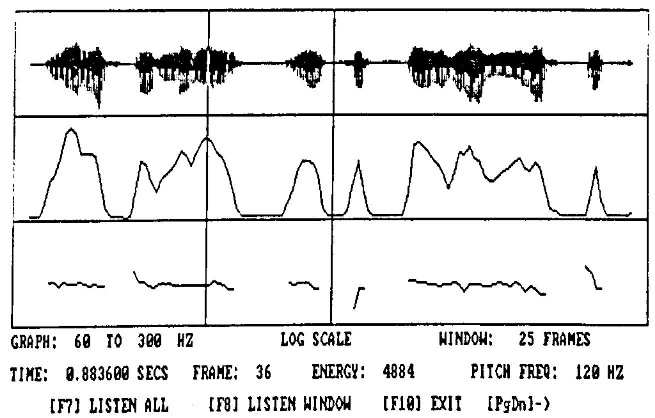
To use the PDB, groups of 2-4 students sit at a "work-station" table around one computer with external speaker and space for writing and reference materials. All words and texts are listed by filename, including language and number, to be typed in when calling up an item. Lists include phonetic and phonemic transcriptions of each item, orthographic representations in the native language if available, and English glosses. The instructor first assigns types of sounds for students to listen for, with lists of files containing those sounds. Initial listening focuses attention at the long-term level, on features of phonation type, voice setting, and dynamics [19]. Some features which can be identified from PDB text samples in this long-term listening exercise include tongue fronting, breathiness, close jaw, backing, spreading, roundedness, nasality, and prominent manners or secondary articulations such as retroflexion, clicks, approximants, frication, affrication, or glottalization. Figure 1 illustrates an MSL EDIT display contrasting recurring clicks of Xhosa (screen A, top) with the acoustic waveform and articulatory characteristics of an Inuktitut text (screen B). The set-up indicates that screens A and B will be heard repeatedly at 1000msec intervals.

Figure 1.
MSL EDIT display of Xhosa (top) and Inuktitut text.



Pitch, amplitude and spectral characteristics can also be calculated and displayed by MSL, adding recognition of visual correlates to the task of becoming familiar with a range of auditory features. This is illustrated in figure 2 where the Inuktitut text has been analyzed to show amplitude (middle screen) and pitch (bottom screen) over time. Read-outs represent values at the position of the left cursor. In figure 2, left and right cursors have been placed to isolate 25 frames of speech. This adjustable window, or the entire waveform, can be monitored using D/A by pressing the function keys indicated in the menu scroll. This capability is also present in MSL EDIT.

Figure 2.
MSL amplitude and pitch display of Inuktitut text.

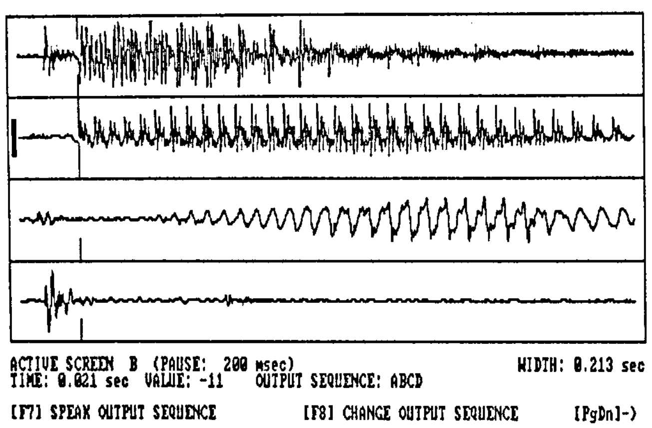


Students locate new sounds, manipulate elements of stored items, arrange them in categories, and create new sets of files that represent the inventory of speech sounds from the phonetic chart to meet the instructional objective of the course. Evaluation of the task considers: (a) number and range of representation of sounds collected, (b) adequacy of each item extracted from surrounding speech to illustrate the sound intended, and (c) organization of items into phonetic categories for presentation. The goal is for students to become active agents in their own learning process while, at the same time, learning the use of instrumental techniques.

Items are collected with task-based instructions: "Find all the words from the following languages that have sound X in them," and then "Find the sounds that the following words have in common" and "Group the following words together according to sounds that they share in common." Once collected, the sounds are studied in detail and gradually assigned phonetic symbols. Sounds isolated in this process are then grouped together in new files representing sets of allophonic variants of the "same" phoneme in a language. Isolated sounds can also be combined in new files represented by the same phonetic symbol, but which have been taken from different languages. Figure 3 illustrates how short samples can be collected, marked and displayed. The cursor in each screen is aligned at 0.021sec to highlight initial consonant differences.

Figure 3.
MSL EDIT display of similar CV sequences.

A	B	C	D
Xhosa [t'ɑ]	Korean [t+ɑ]	Nitinaht [t'ə]	Skagit [t'ə]



If pharyngeal sounds need to be demonstrated, for example, the pharyngealized series of stops, affricates or nasals from Salishan and Wakashan languages, or the pharyngealized series from Arabic, are loaded for auditory contrast and transcription and visual observation of acoustic correlates. Extensive exposure is achieved by having students collect a variety of reflexes of each articulation specified on the phonetic chart, especially for sounds or symbols they find difficult and want to practice.

In another activity format, as a testing or "challenge" procedure, five items are displayed for visual identification. The instructor or a student specifies a sound by phonetic symbol or articulatory label, for a group of students to locate. Cursors can be positioned on the screens to isolate the sound and examine its transitions. If the indicated sound is not present, the item(s) closest to it in articulatory features must be identified.

CONCLUSION

With the development of a Phonetic Data Base, the presentation of speech sound material for phonetic study is facilitated, allowing expedient access to greater amounts of data, and manipulation and organization of speech items in an active learning format. The system also permits the training of language teachers in the use of technological aids for the delivery of speech sound information, in a manner consistent with the precepts of communicative, holistic language learning theory. Research on second language acquisition processes and teaching approaches is integrated with Micro Speech Lab hardware and software for delivery and analysis of speech signals to provide an expedient system for presenting phonetic material for pedagogical purposes. Additional applications of this system include the transmission and sharing of speech data for collaboration in phonetic research.

REFERENCES

[1] C. Dickson, *User's manual for Micro Speech Lab*, Software Research Corporation, 1985.
[2] C. Dickson, *User's manual for the MSL comparison and editing program: MSL EDIT*, Centre for Speech Technology Research Society, 1987.
[3] F. Smith, *Reading without nonsense*, Teachers College Press, 1978.
[4] E. Stevick, *Memory, meaning and method*, Newbury House, 1976.
[5] S. Savignon, *Communicative competence: theory and classroom practice*, Addison-Wesley, 1983.
[6] J. Morley, *Improving aural comprehension*, University of Michigan Press, 1972.
[7] P. Dunkel, Developing listening fluency in L2: Theoretical principles and pedagogical considerations, *Modern Language Journal, 70*, 99-106, 1986.
[8] S. Krashen, T. Terrell, *The natural approach*, Pergamon, 1983.
[9] E. Purcell, R. Suter, Predictors of pronunciation accuracy: A reexamination, *Language Learning, 30*, 271-287, 1980.
[10] R. Wong, Does pronunciation teaching have a place in the communicative classroom? Georgetown University Round Table on Languages and Linguistics, June, 1985.
[11] G. Brown, G. Yule, *Teaching the spoken language*, Cambridge, 1983.
[12] H. Woods, *Syllable stress and unstress*, Canadian Government Publishing Centre, 1978.
[13] J. Esling, R. Wong, Voice quality settings and the teaching of pronunciation, *TESOL Quarterly, 17*, 89-95, 1983.
[14] J. Gilbert, *Clear speech*, Cambridge, 1984.
[15] B. Harmegnies, A. Landercy, Language features in the long-term average spectrum, *Revue de Phonétique Appliquée, 73-74-75*, 69-79, 1985.
[16] H. Dulay, M. Burt, S. Krashen, *Language two*, Oxford, 1982.
[17] S. Krashen, *Second language acquisition and second language learning*, Pergamon, 1981.
[18] T. Terrell, Acquisition in the natural approach: The binding/access framework, *Modern Language Journal, 70*, 213-227, 1986.
[19] D. Abercrombie, *Elements of general phonetics*, Edinburgh University Press, 1967.

Acknowledgements. I am indebted to my co-researchers, Geoff O'Grady, Thom Hess, Barry Carlson, and numerous colleagues and graduate students who are the source of the majority of linguistic material in the Phonetic Data Base. The indispensable cooperation and collaboration of Craig Dickson and Roy Snell of the Centre for Speech Technology Research at the University of Victoria in conceiving and developing microcomputer-based speech processing systems is gratefully acknowledged, as are general university research grants from the Social Sciences and Humanities Research Council of Canada.

GRAFISCHE MODELLIERUNG DER SPRECHBEWEGUNGEN MIT HILFE EINES KLEINCOMPUTERS

GERHART LINDNER

Sektion Rehabilitationspädagogik und Kommunikationswissenschaft, Humboldt-Universität Berlin (DDR) 1040

ZUSAMMENFASSUNG

Die Entwicklung der Computertechnik macht es möglich, die artikulatorischen Bewegungen beim Sprechen auf einem Monitor zu modellieren. Damit sie in der Lehre im Hochschulunterricht eingesetzt werden können, ist ein Programm, das sich mit Hilfe eines Kleincomputers darbieten läßt, von Vorteil. Zu einem Grundmuster des Kopfes werden die Bewegungen der Artikulationsorgane (Lippen, Unterkiefer, Gaumensegel) mit differenzierten Programmen ausgewiesen. Bei Eingabe der Lautfolge werden sie zusammengefügt. Die Zungenlinie wird durch einen besonderen Algorithmus ermittelt, in dem die Laute fixierte Höhepunkte darstellen. Um den Bewegungsablauf auch in den Zwischenphasen wirklichkeitsgetreu abbilden zu können, ist der Vergleich mit Realaufnahmen notwendig. Für die deutsche Sprache wird das Ergebnis der Modellierung mit Röntgen-Zeitlupen-Aufnahmen verglichen, die bereits ausgewertet sind und an denen einige Gesetzmäßigkeiten des koartikulatorischen Bewegungsablaufs ermittelt wurden. Die Optimierung des Programms für die Zungenlinie erfolgt in drei Schritten: 1. Lineare Interpolation, 2. Berücksichtigung des differenzierten Bewegungstempos der Artikulationsorgane, 3. Berücksichtigung der lautübergreifenden Koartikulation.

1. NOTWENDIGKEIT DER MODELLIERUNG ARTIKULATORISCHER BEWEGUNGEN

Das Sprechen ist ein weitgehend automatisierter Prozeß, bei dem eine Vielzahl von dezentralisiert im Organismus gelegenen Organen harmonisch zusammenwirkt. Bei der Anbildung oder der Korrektur des Sprechens ist es notwendig, in diesen automatisierten Prozeß einzugreifen. Um rationell vorgehen zu können, ist es erforderlich, diejenigen Artikulationsbewegungen zu erhalten und zu nutzen, die richtig ausgeführt werden, und gleichzeitig diejenigen Bewegungen zu korrigieren, die fehlerhaft sind. Dazu muß der Pädagoge über differenzierte Vorstellungen und detaillierte Einsichten in den Prozeß der miteinander verflochtenen Bewegungen der Sprechorgane verfügen. MENZERATH hat dieses komplexe Geschehen anschaulich als Sprechbewegungsgefüge bezeichnet./1/

Den Sprechbewegungsablauf darzustellen und zu lehren ist deshalb besonders schwierig, weil

- nur Höhepunkte des als Gesamtablauf über das akustische Klangprodukt kontrollierten Komplexes, die Laute, bewußt werden,
- die Umsetzung der Lautsprache in die Schrift nur diese Höhepunkte nutzt und damit die Orientierung des Kenntniserwerbs in bezug auf Ausschnitte aus dem Gesamtkomplex unterstützt,
- sich die Organe nicht, den isolierten Lauten entsprechend, plötzlich und ruckartig bewegen, sondern im Verlauf eines harmonischen Bewegungsgeschehens die Positionen durchlaufen, die den Lauten entsprechen,
- sich gerade zwischen den als Lauten gekennzeichneten Höhepunkten wichtige Bewegungen einzelner Organe vollziehen,
- das Tempo der Bewegungen der einzelnen Organe unterschiedlich ist, dies aber nur in den akustisch wirksamen Gesamtprozeß eingeht,
- sich die innere Anschauung über das Bewegungsgefüge nicht aus der Selbstbeobachtung gewinnen läßt, da wesentliche Bewegungen der unmittelbaren Beobachtung entzogen sind.

Zum Zweck der Korrektur und der systematischen Anbildung muß der Pädagoge in der Lage sein, den komplexen Bewegungsablauf in seine Einzelheiten aufzulösen. Die Vermittlung dieser Vorstellungen ist ebenfalls schwierig, weil

- die visuelle Beobachtung der Sprechorgane auf die Bewegungen von Lippen, Unterkiefer und Zungenspitze beschränkt ist,
- sich nur wenige Organe taktil kontrollieren lassen,
- die auditive zeitliche Differenzierung wohl zur Erkennung der Laute, nicht aber der Lautübergänge ausreicht,
- sich die Bewegungen, die sich in stimmlosen Perioden des Sprechens vollziehen, weder auditiv noch meßtechnisch über das akustische Signal erfassen lassen.

Mit den phonetischen Anschauungsmitteln, wie sie sich in Lehrbüchern finden, lassen sich zwar Kenntnisse über die lautbezogenen Organpositionen vermitteln./2,3,4/ Diese Kenntnisse betreffen aber nur Höhepunkte des miteinander verflochtenen Bewegungsablaufs und nehmen auf die Veränderungen bei der Koartikulation keinen Bezug. Für die effektive Korrektur des zusammenhängenden und die Anbildung des fließenden Sprechens ist es aber notwendig, daß der Pädagoge weiß, wie sich die Bewegungen der Sprechorgane zwischen den Ausschnitten vollziehen, damit er beim Schüler die notwendigen Bewegungen stimulieren, entwickeln und kontrollieren kann. Denn der Schüler muß beim Sprechen Bewegungen vollziehen, und der Lehrer muß diese bewerten.

Für die Lehre ist es deshalb notwendig, ein anschauliches Modell zu entwickeln, mit dem sich die Bewegungen der Artikulationsorgane beim Sprechen demonstrieren lassen. Dadurch wird der Lernprozeß, der zu anwendungsreifen Vorstellungen führt, abgekürzt und gleichzeitig über das Niveau hinausgeführt, das mit den heutigen Anschauungsmitteln erreichbar ist.

2. VORAUSSETZUNGEN FÜR DEN LÖSUNGSANSATZ

Damit die Aufgabe, die mit Hilfe der Computergrafik lösbar geworden ist, realisiert werden kann, müssen eine Reihe von Vereinfachungen vorgenommen werden. Sie betreffen

- den Verzicht auf die Individualität,
- die Darstellung der Bewegungen in nur zweidimensionaler Abbildung,
- die Beschränkung auf die deutsche Standardaussprache, wobei spätere Erweiterungen auf Sprechfehler, Dialekte oder Fremdsprachen vorgesehen werden,
- die Kontinuität der zeitlichen Auflösung, da die Bewegungen ohnehin auf einem Monitor dargestellt werden.

Damit eine computergrafische Modellierung vollzogen werden kann, werden die Einzelheiten des Kopflängsschnittes nach bewährtem Muster in unbewegliche und bewegliche Organe unterteilt./5/ Die unbeweglichen Teile dienen zum genauen Verfolgen der Bewegungen und werden unverändert beibehalten. Sie sind der konstante Teil des grafischen Programms:

- Oberkiefer mit Zähnen, Ansatz zum Nasenraum, hintere Rachenwand, Kehlkopf mit Stimmlippen in einer Mittelstellung.

Die artikulatorischen Bewegungen während des Sprechens werden von folgenden Organen dargestellt:

- Unterkiefer mit Zähnen und Kinnlinie, Oberlippe, Unterlippe, Gaumensegel und Zungenlinie.

Die Einstellungen und Bewegungen der Stimmlippen sind in der zweidimensionalen Darstellung des Kopflängsschnittes nicht darstellbar. Für eine Weiterentwicklung ist geplant, sie in einer anderen Abbildungsebene (Draufsicht) in einer Ecke des Bildes einzublenden. Die aktuelle Lautfolge der modellierten Bewegungen wird in phonetischer Umschrift dargestellt.

3. PRINZIPIELLE SCHWIERIGKEITEN EINER ARTIKULATORISCHEN BEWEGUNGSSYNTHESE

Die Modellierung artikulatorischer Abläufe ist deshalb schwierig, weil es sich um die Darstellung von Einzelheiten handelt, die von der Phonetik bisher nicht oder kaum erarbeitet worden sind, da sie praktisch nicht gebraucht wurden. Sowohl bei der Anbildung und Korrektur des Sprechens als auch im Fremdsprachenunterricht können sich Lehrer und Schüler auf voll eingeübte Bewegungsautomatismen stützen. Auch bei einem völlig Stummen sind Bewegungsvollzüge der Organe, die zum Sprechen notwendig sind, für den rein emotionalen Ausdruck und zur Nahrungsaufnahme in einer bestimmten Weise eingespielt. Diese Bewegungen müssen im Verlauf des Lernprozesses umgestellt und anders koordiniert werden. Bei der Ausführung dieser natürlichen Bewegungen der Sprechorgane sind die einzelnen Teile, die im Modell separat dargestellt werden müssen, zwangsläufig miteinander verbunden und in ihrem Bewegungsinventar aufeinander abgestimmt.

Für die Modellierung der Bewegungen fallen diese Bedingungen weg, die das Bewegungsinventar einschränken. Außerdem entfällt die Möglichkeit, das auf dem Monitor Modellierte durch den akustischen Effekt zu kontrollieren. Für die Modellierung gibt es bisher keine Vorgaben für Grenzbedingungen der Einstellungen oder Bewegungen einzelner Organe und für deren Zusammenwirken; denn solche Angaben werden für den Unterrichtsprozeß nicht gebraucht. Deshalb steht die Modellierung von artikulatorischen Bewegungen heute vor Schwierigkeiten, die denen gleichen, die die akustische Sprachsynthese zu Beginn ihrer Arbeiten überwinden mußte. Das System für die Sprechbewegungssynthese ist ein offenes System. Mit ihm sind prinzipiell alle bildlichen Darstellungen der Sprechorgane realisierbar, auch solche, die auf den ersten Blick als unsinnig erkannt werden (z. B. wenn die Zunge aus zwei Teilen besteht). Die Schwierigkeit besteht darin, daß die Darstellung exakt der Wirklichkeit entspricht, damit dem Schüler wirklichkeitsgetreue und anwendbare Vorstellungen von Bewegungsvollzügen vermittelt werden.

4. VERFAHREN DES LÖSUNGSWEGES

Damit ein Gesamtbild modelliert werden kann, muß es nach bewährten Grundsätzen, die auch für bewegte Trickzeichnungen angewendet werden, in einzelne Teile zerlegt werden. Als Teilbilder werden benutzt:

- Das Grundmuster des Kopflängsschnittes. Es wurde nach anatomischer Vorgabe entworfen./6/
- Die verschiedenen Einstellungen des Unterkiefers. Zwischen maximaler Weite und minimaler Enge wurden weitere 7 Stellungen vorgesehen. Als 10. Position ist eine retrahierte Stellung geplant, um die labialen Engelaute darzustellen.

- Die aktiven Einstellungen von Ober- und Unterlippe. Dabei setzt die Oberlippe an feststehenden Punkten des Grundmusters an. Die Bewegungen der Unterlippe sind von der Position des Unterkiefers abhängig. Obwohl bekannt ist, daß in bestimmten artikulatorischen Abläufen die aktiven Bewegungen von Ober- und Unterlippe nicht übereinstimmen, wurden sie auf der Grundlage einer gemeinsamen Aktivität modelliert.
- Die verschiedenen Stellungen des Gaumensegels, dessen Darstellung an Festpunkten des Grundmusters ansetzt.
- Für die Erweiterung des gegenwärtigen Programms ist auch die Modellierung der Glottiseinstellungen vorgesehen. Dabei ist zu berücksichtigen, daß neben der Stimmstellung auch die Einatmungs-, Hauch-, Flüster- und Ruhestellung zu sehen sind.

Die Bewegungen von Unterkiefer, Ober- und Unterlippe, Gaumensegel und auch der Glottis lassen sich mit Hilfe eines relativ kleinen Programms modellieren, da sich diese Bewegungen immer in der gleichen Weise wiederholen und die Variationsmöglichkeiten gering sind.

Größere Schwierigkeiten bereitet die Modellierung der Zungenlinie, da sie einerseits von den Bewegungen des Unterkiefers abhängig ist und andererseits in mindestens zwei Hauptabschnitte (Zungenspitze und -rücken) aufgeteilt werden muß.

Um diese komplizierte Bewegung mit vielen Übergangsmöglichkeiten modellieren zu können, wurde ein spezielles Koordinatensystem entwickelt.

Es hat sich als ausreichend erwiesen, die Zungenlinie mit 15 Punkten festzulegen und die Zwischenräume durch gerade Strecken miteinander zu verbinden. Diese Punkte liegen an den Stellen maximaler Krümmung. Da bei den Vokalen der gesamte Hohlraum des Ansatzrohres an der Klangbildung beteiligt ist wurde entschieden, alle Werte der Zungenlinie fest zu speichern. Bei den Konsonanten wurde eine Differenzierung vorgenommen, indem jene Gebiete der Zungenlinie, die unbedingt eingenommen werden müssen, von solchen unterschieden werden, die bei der Lautbildung koartikulatorisch veränderlich sind.

Das Programm wurde von Anfang an so konzipiert, daß es für weitere Laute ergänzungsfähig ist und daß es in einem späteren Stadium mit dem Sprachsynthesator ROSY 4201 /7/ zusammengeschaltet werden kann, so daß dann die Möglichkeit gegeben ist, die visuelle Synthese mit der akustischen zu kombinieren.

Aus den Teilbewegungen von Unterkiefer, Lippen, Gaumensegel und Zungenlinie läßt sich die beim Sprechen einer beliebigen Lautfolge notwendige Bewegung der Artikulationsorgane als Ganzes modellieren. Dabei reicht es aus, die Bewegungen, die sich in dem Zeitraum von 20 ms vollziehen, zu einem Bild zusammenzufassen. Eine feinere Unterteilung ist weder nötig noch möglich, da auf dem Monitor pro Sekunde nur 50 Halbbilder gezeigt werden.

Für den Zeittakt wurden die folgenden Vorgaben konzipiert, um die Wirklichkeitstreue der Modellierung zu überprüfen. Die Vorgaben für den Zeittakt wurden aus der Literatur entnommen /8/ und grob vereinfacht. Die Zeittakteinheit (ZTE) wurde als 20 ms festgelegt. Es beträgt die Dauer

- eines kurzen Vokals 4 ZTE
- eines langen Vokals 7 ZTE
- eines Diphthongs 7 ZTE
- des Murmelvokals 3 ZTE
- eines Konsonanten 5 ZTE

Diese Zeitvorgaben werden in Abhängigkeit von der Akzentuierung modifiziert:

- Zuschlag zu einer betonten Silbe 3 ZTE
- Kürzung bei einer unbetonten Silbe 2 ZTE

Der Vergleich mit dem real produzierten Sprachmaterial wurde an dem gleichen Testsatz vollzogen, der röntgenkinematografisch aufgenommen und ausgewertet worden war. /9/ Dieser Testsatz, der die häufigsten deutschen Lautfolgen enthält, lautet: "Wie denn, ist das einer der Steine, die ich im Winter anderwärts gefunden habe?" Die Zeit für die Modellierung ergibt nach den oben angegebenen Zeittaktwerten eine Dauer von 5,36 s und liegt damit innerhalb der individuellen Variationsbreite.

5. STANDORTBESTIMMUNG FÜR DIE REALISIERUNG

Wenn man davon ausgeht, daß die Laute Höhepunkte des artikulatorischen Geschehens sind, dann ist es möglich, die für die Laute entwickelten Gesamtbilder als die Ausgangspunkte für die Modellierung der Übergänge zu nutzen. Es wird zunächst davon abgesehen, daß auch die Lautpositionen im zusammenhängenden Sprechen veränderlich sind.

Es ist bekannt, daß vor allem die Konsonanten, aber auch die Vokale, koartikulatorischen und akzentabhängigen Veränderungen unterliegen, was sich besonders in der Realisierung unbetonter Silben der Rede ausdrückt. Sie sind für das Russische besser aufgearbeitet als für das Deutsche. /10/

Für das Deutsche liegen sie weder in algorithmisch nutzbarer Form noch in verwertbaren Abbildungsunterlagen vor. Deshalb wird bei der Modellierung zunächst an unveränderlichen Lautbildern festgehalten. Sie stellen im Modell die Phasen dar, die die Artikulationsorgane bei der Aussprache einnehmen oder wenigstens durchlaufen. Bei den langen Anteilen eines Lautes wird dann die für den Laut vorgegebene Position eine Anzahl von Phasen gleichgehalten. Bei den kurzen Lauten wird diese Phase zwar als Zielposition für die Modellierung zum Höhepunkt hin verwendet, aber nach deren Erreichen sofort als Ausgangsposition für die Ansteuerung zum nächsten Laut verwendet. Der Betrachter erlebt dann diese Position eines kurzen Lautes nur als Durchgangsphase.

Als einfachste Möglichkeit, die Übergänge zwischen den Lauten zu realisieren, bietet sich die lineare Interpolation an. Dieses Verfahren ist rechnerisch unkompliziert. Damit die Veränderungen, die sich innerhalb einer Lautfolge vollziehen, gut perzipierbar sind, werden, ehe sich das Bild zu verändern beginnt, die auszuführenden Veränderungen als Vektoren abgebildet. (Abb. 1)

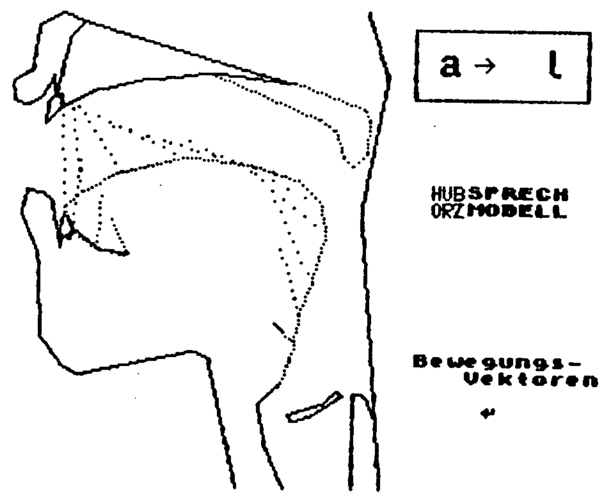


Abb. 1: Computerausdruck eines Positionsbildes mit den Bewegungsvektoren zum Folgelaut.

Die zweite Möglichkeit ist, die unterschiedliche Bewegungsgeschwindigkeit der Organe zu berücksichtigen. Zungenspitze und Lippen sind schnell-, Unterkiefer und Gaumensegel sind langsambewegliche Organe. Mit diesem Algorithmus, der vor allem die hohe Beweglichkeit der Zungenspitze berücksichtigt, wurde zeichnerisch die Lautfolge "der Steine" modelliert. Dies ist ein Ausschnitt aus dem Testsatz und liegt als röntgenkinematografische Analyse von verschiedenen Versuchspersonen vor. Der Vergleich zwischen Modellierung und Realaufnahmen erfolgte mittels des Motogramms. /11/ Der Vergleich zeigt gute Übereinstimmung zwischen Modellierung und Realaufnahmen. Die Modellierung liegt innerhalb der individuellen Variationsbreite.

Die dritte Stufe der Annäherung an die Wirklichkeit ist noch perspektivisch. Sie besteht in der Berücksichtigung der koartikulatorischen Bedingungen, wobei berücksichtigt werden muß, daß die Konsonanten den stärkeren koartikulatorischen Veränderungen ausgesetzt sind.

6. MIT DEM PROGRAMM ERZIELTE ERGEBNISSE

Das Programm wurde mit dem Ziel ausgearbeitet, daß es als Lehr- und Unterrichtsmittel verwendet werden kann. Deshalb wurde es für den an allen Universitäten und Hochschulen der DDR vorhandenen Kleincomputer KC 85/2 ausgearbeitet. Damit ist bewiesen, daß ein so umfangreiches und kompliziertes Programm, wie es für die Bewegungssynthese notwendig wird, mit einem Kleincomputer realisierbar ist. Für einen Teil der Studenten wurde dadurch, daß das Programm in den planmäßigen Lehrprozeß des Phonetikunterrichts einbezogen wurde, die Einsicht in die Dynamik der Physiologie des Sprechens erheblich verbessert, die eigene Beschäftigung mit artikulatorischen Bewegungsvorgängen wesentlich stimuliert.

Die Mängel, die das Programm heute noch aufweist, zeigen, daß noch intensive Forschungen durchgeführt werden müssen, um unsere Kenntnisse über die Physiologie artikulatorischer Bewegungsvollzüge zu erhöhen. Diese erst bilden die Voraussetzung für eine in allen Belangen wirklichkeitsgetreue Modellierung. Es ist anzunehmen, daß die prinzipielle Möglichkeit, Sprechbewegungen computergrafisch zu modellieren, der Phonetik einen starken Impuls für diesbezügliche Forschungen verleiht.

7. LITERATUR

/1/ MENZERATH, P. u. A. DE LACERDA: Koartikulation, Steuerung und Lautabgrenzung. Bonn 1934
/2/ WÄNGLER, H.H.: Atlas deutscher Sprachlaute. Berlin, 7. Aufl. 1981
/3/ Autorenkollektiv (Leitung: H. STÖTZER): Großes Wörterbuch der deutschen Aussprache. Leipzig 1982
/4/ FANT, G.: Phonetik und Sprachforschung. In: Handbuch der Stimm- und Sprachheilkunde (LUCHSINGER/ARNOLD). Wien/New York 1970, S. 274
/5/ LINDNER, G.: Grundlagen und Anwendung der Phonetik. Berlin 1981, S. 89 ff.
/6/ RAUBER/KOPSCH: Lehrbuch und Atlas der Anatomie des Menschen. Bd. II, Leipzig 1955, S. 66
/7/ MEHNERT, D.: Analyse und Synthese suprasegmentaler Intonationsstrukturen, ein Beitrag zur Optimierung technischer Sprachkommunikationssysteme. Diss. (B), Techn. Univ. Dresden, 1985, S. 98 ff.
/8/ LAZICIUS, J.: Lehrbuch der Phonetik. Berlin 1961, S. 120 ff.
/9/ LINDNER, G.: Der Sprechbewegungsablauf. Eine phonetische Studie des Deutschen. Berlin 1975, S. 97 ff.
/10/ GABKA, K.: Einführung in das Studium der russischen Sprache. Bd. I Phonetik und Phonologie (E. WIEDE). Leipzig 1 1974, S. 85 - 95
/11/ LINDNER, G.: Das Motogramm - ein Mittel zur Veranschaulichung artikulatorischer Verläufe. ZPSK 30 (1977), S.535-543

FOR AN UP-TO-DATE VISUAL REPRESENTATION OF SPEECH

Raymond THOMAS, Laboratoire de Langues, Fac.Sciences de Luminy, 13288 MARSEILLE Cedex9, France
Serge BAGNOLI, Fac. Sciences St-Charles, U. d'Aix-Marseille I, 13331 MARSEILLE Cedex 3, France
Jacques GENIN, DAII-SEDE, 38-40, rue du Général Leclerc, 92131 ISSY LES MOULINEAUX, France
Robert H. GREEN, Dpt of Psychology, Ann Arbor University, Michigan 48100, USA
Hubert GREVEN, Fac. de Lettres et Sc. H., Université de Rouen, 76130 MONT SAINT AIGNAN, France
Nils-Olof JØNSSON, Språkpedagogisca Forsknings AB, Box 63, S-230 10 SKANØR, Sweden
Antony McKENNA, Laboratoire de Langues, Fac.Sciences de Luminy, 13288 MARSEILLE Cedex 9, France
William WEISS, Dpt of Theatre, University of Ottawa, OTTAWA, Ontario, CANADA K1N 6N5

RESUME

Une meilleure information sur les connaissances acquisesen acoustique de la parole pourrait être immédiatement disponible pour tous grâce aux techniques informatiques. Nous donnons un exemple de la forme qu'elle pourrait adopter et prions membres et organisateurs du Congrès de préparer l'établissement d'un organisme représentatif qui poursuivrait la réalisation d'une image visuelle, probablement syllabique, de la parole, informant ainsi d'une façon satisfaisante de l'acquis scientifique en ce domaine.

AIMS

At recent meetings of acousticians, I have pointed to the predicament in which we, who teach foreign languages, find ourselves: we must know what speech habits the student should change in order to speak another language well, but the scientific information about pronunciation that we need for that purpose, is too often lacking, or practically unavailable, or uncertain. At the 1986 International Congress on Acoustics at Toronto, I was asked to give a model of what we precisely want. As may be expected, that question can be answered only in part, and that's why the drawings we'll show to give a clear answer, will be somewhat like dreams, made of both true and doubtful elements. Nevertheless, as will be seen, they should help in stating more precisely what elementary problems still plague our general knowledge of speech, and suggest steps to be taken in order that better information on this subject be made readily available to all.

WHAT PARTS ARE BEST DESCRIBED?

Though speech statements are unpredictable, their parts are limited in diversity, and thus susceptible of scientific description. The most obvious elementary parts in speech chains are the syllables, but these have long been divided into vowels and consonants, and lately these smaller parts, in their turn, were analysed into smaller elements variously described as phonemes, features, indices, etc. The first problem therefore is: what elementary parts of pronunciation should be selected for a practical, realistic, and up-to-date visual description of speech?

Three Reasons for Choosing Syllables.

Tentatively in these drawings, we have chosen syllables as basic units, and for a long time past, the three following reasons could have been given for this choice. First, they are the only divisions that are immediately perceptible to all language users. Second, we may therefore say that syllables are the basic units used by our minds to process speech if, for the sake of brevity, you forgive us for using such simple terms. Thirdly, the finer divisions vary according to the syllables in which they are to be found, and, for a scientific description, these variations nullify the fundamental advantage of such smaller elements, i.e. that of being less numerous than syllables.

Why Finer Divisions Are Unsatisfactory.

If vowels and consonants, and all the more, phonemes, features, etc., vary according to the other components with which they are to be found, two attitudes may be adopted. First, if we want to give a realistic visual representation of speech with vowels and consonants, we must distinguish between the various sounds that each of these may represent. This recently led to the International Phonetic Alphabet, and we know that such notations cannot be satisfactory because too many diacritic signs make them impracticable, even though only the more important variations are noted.

A second attitude is to forgo writing as an unequivocal system of signs; this is what we generally do and, particularly for the study of our Western languages, it is one of the basic obstacles students have to overcome. But, more fundamentally, such an equivocal use of vowels and consonants cannot be a satisfactory basis for a scientific description of language sounds since what is then spoken of, in one instance, may not be the same object as what is described in another, without our being aware of this change in the object. Many problems of present day acoustical research should be ascribed to this fundamental cause.

The Decisive Reason for Choosing the Syllables.

But the practical reason which motivated our choice of the syllables as the basic units to be described in the speech chain, is pedagogical. For over twenty years now, teaching experience has consistently shown that a predominant attention to syllables,

rather than to the individual vowels and consonants, is a decisive factor in helping English and French speakers to use the other tongue without those irrepressible defects that mark our two communities as particularly unsuccessful among European students of foreign languages. This pedagogical advantage of stressing attention to syllables does not seem to be restricted to French and English, but to apply it to other languages requires special studies for each of these, and this was why we first asked acousticians for an improved general description of language sounds. The benefits for the millions of language students the world over, suggest the importance of the pursuit, and its interest is not limited to pedagogy.

A New Reason for this Choice.

Lastly, finer analyses of speech sounds have recently shown that vowels and consonants are not always separated in the time sequence as we formerly thought and as is recorded in our spellings. On the contrary, syllables follow strictly the sequential character of speech. At bottom, this may be but another aspect of the first two reasons we mentioned, but it is a most compelling argument when considering visual representation.

HOW COULD SUCH DESCRIPTIONS BE ACCOMPLISHED?

This question requires two different answers. As we were invited to do, we must give a precise view of what we would find in such new descriptions, and this we shall do with our drawings. But a more practical answer is also needed: showing what should, or could, be doneis in itself ineffectual; we must also say how the necessary work can be accomplished. This need for a better description of languages is not just theoretical; congresses such as this are occasions to start a process leading to some organization for a cooperation among acousticians who can enlighten us in their various fields.

The first step we mentioned: switching more scientists' attention to syllables, is too heavy a task for anyone, and therefore it requires the action of such a representative body. And in the course of our description of a desirable new visual representation of speech, the need for such a research organization will again be found in several convincing ways.

A FEASIBLE SYLLABIC VISUAL REPRESENTATION

Though it cannot be done in this black and white paper, we would take advantage of the widening use of color and information processing. We would have three separate drawings for production, transmission, and decoding of syllables. In these drawings, the common parameter of duration will be shown on the horizontal or "x" axes of our two-dimensional models. We have assigned different colors to each of the drawings by specifying the lengths of the corresponding light-waves, as will be seen in the fourth view that can be shown only with colors. In black and white, the superposition of lines and areas would be confusing.

Light-wave lengths between 7200 and 5600 Angströms may mark the range of colors to be used in figure 1; for Fig. 2, the choice would be more limited (5600 to 5000 A.) since the diversity in the airwaves pressures is not so large as for the phenomena described in Fig. 1 and 3. For this third figure, the corresponding lengths would extend from 5000 to the end of the specter.

Using light-wave lengths as practical units may be deemed artificial, and since these drawings will be interpreted with the help of computer processing, the specifications of colors that computers manufacturers provide, could be more conveniently used. Hopefully, our research organization could work for standardization of these color specifications.

Further Notes on the Proposed Representation of Syllables.

The various intensities, or degrees, in the recorded phenomena, could be shown by the intensity of the colors used. Within each of the three groups of recorded phenomena, related facts or events could be pictured in the same colors. Some type of color correspondance could also be specified for related parts in the three drawings, whenever such relations are generally recognized.

Our drawings show individual production of precisely determined syllables, and not more or less general models of syllables, since "norms" and "averages" are not yet established for a sufficient classification among them. Only an important research organization could usefully set standards or "averages" to obtain such classifications.

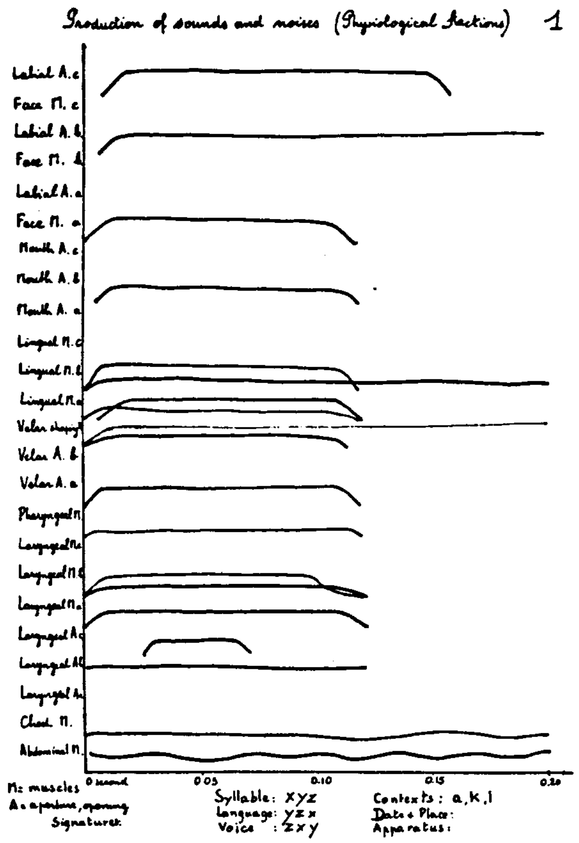
"Model" or "average" syllables could thus be determined, but several types of occurrences must be separately represented. To give just one instance of this unavoidable multiplicity, the duration of accented syllables cannot usually be assimilated to that of unaccented syllables. In our use of this complex information, a limited number of properly selected syllables would be sufficient for an efficient study of the sound characteristics of a given foreign language. On the other hand, storing the large volume of the pertinent information on all the syllables in any language, is not a major problem for computers. Similarly, the calculations needed to obtain "average" or "normal" models from a sufficient number of individual performances does not seem to be unusually large either.

Since such "models" are not available now, our present drawings show the performances by a "normal" speaker or hearer in a given language for what we deem a representative occurrence of the chosen syllables.

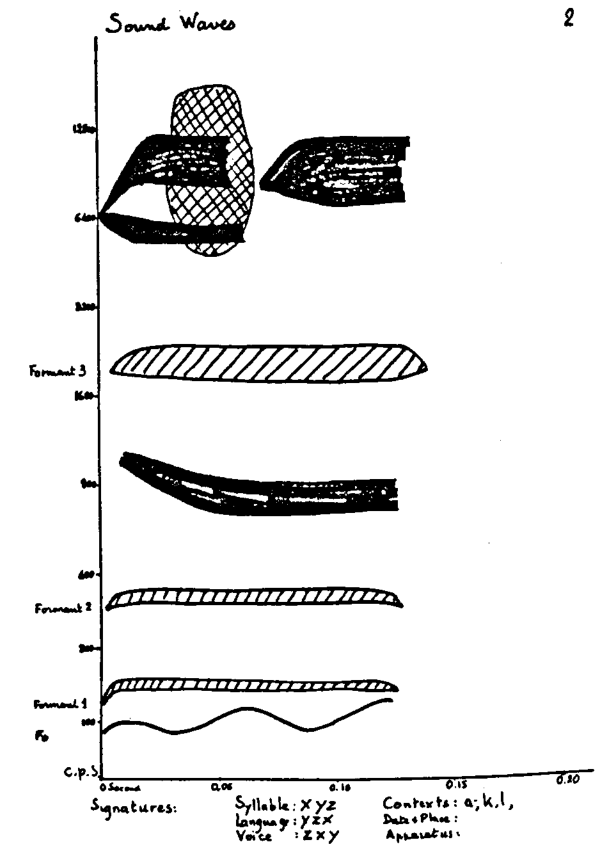
We must also note that, similarly to the establishment of "norms" of syllable occurrences, the decisions on many points in our drawings require a knowledge and an authority that can rest only with a representative research organization.

In our drawings, we have purposely included some obviously unusual opinions in order to underline both that these pictures are intended just to show one possible kind of up-to-date visual representation and that our generally accepted opinions may have to be modified, for instance as a result of a wider study of speech in the various communities.

FIGURE 1 : SPEECH PRODUCTION



F.2 "TRANSMISSION OF SPEECH" OR"AIR PRESSURE WAVES"



In this, and in the third pictures, we should imagine as many "x", or horizontal, axes, as there are phenomena, or groups of phenomena, indicated in the left margins. In these drawings, intensities or degrees are often marked by curves, and therefore various intensities of colors are not needed here; so these could be used for information that is not now available about these phenomena. Though we have hinted at much of present day knowledge, we beg to be forgiven for insisting that these drawings are not meant to be complete, but merely to show how knowledge about language sounds may be accurately recorded in visual representations.

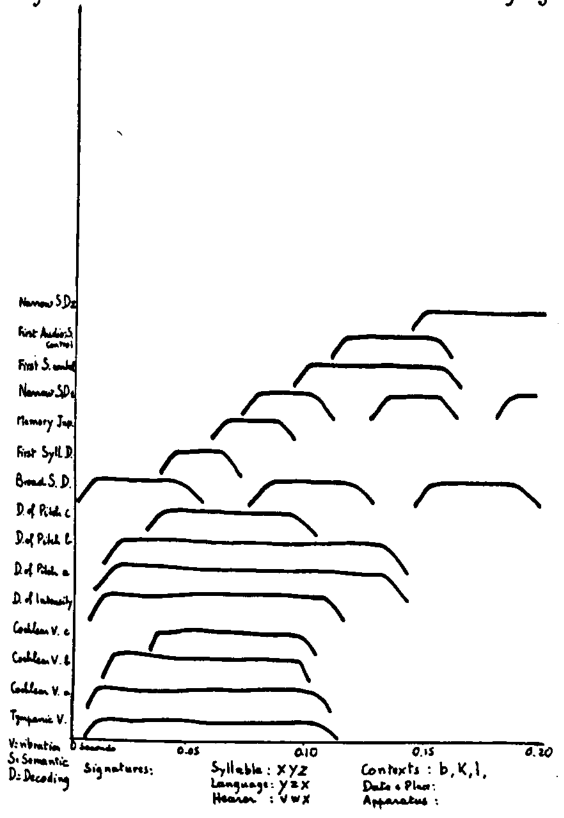
Perhaps we should have grouped together muscular actions in the higher or lower part, and the positions of the various organs of speech in the other part of the figure. The aspect of the picture would change much according to such options. There are at least two often conflicting rules to be followed in all our drawings: 1/in the conventions we adopt, we should try to keep as near to actual facts as we can; 2/the drawings should be as expressive as possible. In this, and still more in the fourth totalizing picture, various factors will be mixed that cannot readily be isolated by an untrained user. However, though in a different way, this is also true for the sounds of real speech. So long as the same conventions are strictly adhered to for all representations, their artificial character is not to be unduly regretted.

Since neither the voice, nor the language, and not even the syllable are specified, this drawing is most arbitrary, and the sounds of speech it partly represents, may cause more surprise than in the first figure. In the fundamental frequency, we purposely marked a variation which is unusual in our European languages, in order to stress the fact that most of the five thousand, or so, languages have not yet been scientifically studied, not even briefly. Therefore, what we consideras basic and general rules of the production, transmission, and decoding of speech, could be less universal and necessary than we hold.

Fig. 3 DECODING THE NATURE OF THE TOTAL SOUND.

Picture on top of next column. For a long time, this third view should remain the least complete of the three. Recognition of pitch is not achieved in one way only, as could be inferred from our drawing; but that is not the only point for which more knowledge could already be noted. Still, in our picture we have shown our opinion that it is not realistic to describe syllables produced outside meaningful speech, without due caution. To be sure, not everybody would agree that, meaning being essential to speech, an analysis that unguardedly excludes the meaningful aspect of language, runs a risk of being pointless or wrong. Perhaps will it still take many years before this opinion is examined by influential persons, and yet we have noted it in this third figure,

Fig.3 Aural reception and semantic(?) decoding Fig.3



to show that such representations of speech could be used to other ends than mere information.

To defend the validity of this opinion however, perhaps we might note that, even in a language like French, in which syllables are relatively unaffected by the meaning they convey, there are still important variations due to their specific role in a sentence. For instance, the syllable "in" which often means"negation" in some way or other, sounds quite differently in words such as "inexact","innombrable" and "impossible", and still more unexpectedly in "innovation" and "innové". Even more directly in keeping with our third drawing, the syllable "ma" may be more easily identified as a part of a word like "matériel" than if it is pronounced separately, as in contrast with "pa" or "ba", "fa", etc.

Semantic decoding cannot be overlooked therefore, without unpredictable consequences for an objective study of speech. We have even noted that, for the same speech fact, contexts generally differ for the speaker and the hearer. We have suggested only minor differences in this case, but in our view, this point cannot be overlooked.

THE NEED FOR A REPRESENTATIVE BODY

These facts are not usually taken into account, and more generally speaking, points that require discussion are many. Only a representative body can decide on these controversial issues. Realistic drawings should be aimed at, and the relative serenity of a representative institution is most desirable to achieve this. And perhaps, for the third drawing, should we have stressed even more than we did, that our attitude about semantics was marked, not with a view to have it accepted, but especially to show the inclusive character that these new visual representations of speech could have. A limited organization, not to speek of an isolated individual, cannot be responsible for such a vast endeavor.

CONCLUSION

Especially in their totalising pictures, which cannot be given without colors, our successors may bring a measure of beauty that must now remain a conjecture. But, as we have beautiful scripts and prints, perhaps our conventions will be modified in the future with this preoccupation in mind.

For now, let's limit ourselves to voicing a wish that is probably in each of us: to have an easy and accurate means of access to all the findings which our colleagues will have gathered on our subject.

But, to be practical in this conclusion, our duty now is to beg you, members and organizers of this meeting, not to leave it to a hypothetical future to take steps towards establishing a representative authority that can change our present dream into a not too far distant reality.

BIBLIOGRAPHY

BERNARD J. Comprendre et organiser le traitement automatique de l'information Dunod Paris 1967.

CHOMSKY N. Language and Mind Harcourt NY 1968.

DELATTRE P. "De la hiérarchie des indices acoustiques pour la perception de la parole" Proc. 5th IC Phon. Sciences 1962 244-251

" Principes de phonétique française, Middlebury College, Middlebury Vermont 2è éd. 1951

EIMAS P.D. "The Perception of Speech in Early Infancy" Scientific American Jan. 1985

FOUCHE P. Traité de prononciation française Paris, Klincksieck 1956

GENIN J. Contribution à l'étude des systèmes de prothèse auditive sensorielle Ms thèse BU Grenoble'87

GREVEN H. Elements of English Phonology Paris Puf '72

HAGEGE C La structure des langues Paris Puf 1982

LAVER J. The Phonetic Description of Voice Quality Cambridge UP 1980

LENAT D. B."Computer Software for Intelligent systems" Scientific American September 1964

POLANYI M. "Sense-Giving and Sense-Reading" Philosophy Macmillan Oct. 1967

SAUSSURE F. (de) Cours de linguistique générale 1916

THOMAS R. Evaluation de l'information en langues naturelles Manusc. Thèse BU Rouen 1983

" "Teaching Foreign Pronunciations" Abstract in JASA Suppl. 1 Vol 76 Fall 1984 Minneapolis

" "Recherches pour l'enseignement des prononciations étrangères" JEP ENST Paris 1985

" Pour décrire efficacement les phonétismes" 12 ICA Toronto 1986.

Time-Life Les images électroniques coll. Le Monde des ordinateurs, Amsterdam Déc. 1986

WINOGRAD T. "Computer Software for working with language" Scientific American September 1984.

PERCEPTION OF PHONETIC FEATURES IN SPEECH CODERS FOR MOBILE COMMUNICATIONS

Maurizio COPPERI and Franco PEROSINO

CSELT - Centro Studi e Laboratori Telecomunicazioni -
Via Reiss Romoli 274 - 10148 Torino (Italy)

ABSTRACT

This paper deals with the simulation in real time and the formal subjective evaluation of two low bit-rate speech coders, viz. LPC and RELP, in a mobile satellite system.

The effects of channel impairments, such as multipath fading and shadowing, on intelligibility scores is evaluated by means of the Diagnostic Rhyme Test. The subjective data have been examined to pinpoint the fidelity with which distinctive features and specific phonetic cues are transmitted. Results show that a RELP coder at 9.6 kbit/s, incorporating an error protection scheme, provides a moderately good quality, while the 2.4 kbit/s LPC vocoder yields a quality that is not felt to be commercially acceptable.

1. INTRODUCTION

In our society, mobile communications have become a need for people and a major objective of research. The perception of coded speech under real-world (noisy) transmission conditions is an important aspect of this area, with several implications into the reliability and quality of existing and/or new services (e.g. mobile satellite communications, cellular mobile telephony, etc.) and into the design of efficient and robust speech coding systems.

In this work, two speech digitizers, that is a Residual Excited Linear Predictive (RELP) coder at 7.2/9.6 kbit/s [1], and a Linear Predictive Coder (LPC) at 2.4 kbit/s [2], have been simulated and assessed through the Diagnostic Rhyme Test (DRT) [3]. Useful diagnostic information on specific quality degradations can also be obtained using phonetically constrained sentences [4], instead of rhyming word pairs. Both RELP and LPC algorithms have been used in mobile communications systems [5,6].

Our major objective was the determination of trade-off relationships between speech intelligibility and channel-capacity requirements in mobile satellite systems. In this context, the main constraints to be faced are due to the available bandwidth and transmitter power of satellite and terminals. The bit rate reduction offered by efficient speech digitizers represents an economic incentive in expanding satellite communications, but the attainable subjective quality is of concern if the service should be extended from professional users to the general public.

More specifically, the coders must be capable of providing acceptable quality also in the presence of multipath propagation and inherent signal fading. This degradation, which typically causes burst errors on the transmission link, can be mitigated by the use of error control techniques. The issue of error control was investigated by exploiting four channel models of increasing complexity to choose the optimal method. Results presented throughout this paper have been obtained using the McCullough model [7], which is characterized by 4 independent parameters and can be used to generate sequences that are similar to real-life error sequences. In particular, in this study we consider two examples of bursty channel environments, the former (channel No. 1) typical of land mobile communications for open area into rural, the latter (channel No. 2) including multipath fading. A brief description of the coders is given in section 2. The DRT structure is described in section 3, and the diagnostic scores are discussed in section 4.

2. CODING TECHNIQUES

Both speech compression algorithms and channel models were simulated in real time on an array processor FPS-120 B connected to a VAX 11/785. These coding systems run in half duplex and use a specific audio processing front end with 14-bit A/D D/A converters. The input speech is band-limited to 200-3400 Hz and sampled at 8 kHz. An automatic gain control circuit permits a suitable reduction of the input dynamic range.

The 2.4 kbit/s LPC is based on a 10th order autocorrelation analysis performed every 22.5 ms, an AMDF pitch extractor with median smoother, and a voiced/unvoiced detector driven by the energy ratio between high and low frequency regions.

The 7.2/9.6 kbit/s RELP coder used in this study performs an 8th order autocorrelation analysis over frames of 25 ms in duration, with Hamming windowing of 37.5 ms. After inverse filtering, a 1000 Hz low-frequency portion (baseband) of the residual signal is quantized and transmitted. The regeneration of the full band excitation signal is performed at the receiver using the spectral folding method [8]. The 9.6 kbit/s RELP incorporates an error protection scheme based upon the combination of bit interleaving and bit protection with error correcting codes. The former mechanism is aimed at splitting a long error burst into several shorter bursts (ideally, into isolated errors), thus allowing, through a sort of "divide-and-conquer" strategy, easier protection of the most important parameters in the data frame. The latter mechanism protects the reflection coefficients k(1) through k(4) using four (15,5) BCH codes, and the r.m.s. value of each frame using a (12,4) code. The first code can correct up to 3 errors, whereas the second can correct 1 or 2 errors. Residual samples are left to the channel mercy. Overall, the frame format of the 9.6 kbit/s RELP consists of 190 bits of speech information, 48 bits of error protection and 2 bits for synchronization. Both the LPC vocoder and the 7.2 kbit/s RELP do not exploit error protection.

3. PROCEDURE

A set of four DRT lists was selected for the experiment. Each list contains 116 pairs of English isolated words, read by native American speakers (2 lists read by males, and 2 read by females). These lists were recorded in a quiet environment using an Altec 659A dynamic microphone without a puff screen.

Six different circuit conditions have been examined, combining the three coding bit-rates with two typical channels, as stated in the introduction. Output signals of the processed stimuli have been recorded on analog tapes and then used for the subjective test. Eight listeners took part in the DRT sessions, that were conducted at the Dynastat Inc. (Austin,Texas) in-house speech evaluation facility.

3.1 Structure of the DRT

The DRT of Voiers [3] is based on discrimination between two rhyming monosyllabic words that differ for the initial consonant. The listener's task is simply to indicate which word has been presented. Word pairs are chosen so that initial consonants differ for only one distinctive feature according to the taxonomy shown in Table 1, in which the sign + means positive (present) state of the feature, the sign means negative (absent) state, and the circle means "doesn't apply". Table 2 shows an example of stimulus words used in DRT.

DRT data can be scored in different ways, according to the investigator's interest. In our work, we want to focus not only on the six major features ,i.e. voicing, nasality, sustention, sibilation, graveness and compactness, which are recognized as essential to phonemic distinction for English, but also on scores for the apprehensibility of a given feature, e.g. sustention, in voiced and unvoiced phonemes, or voicing in frictional and nonfrictional phonemes. That is because a finer examination may often pinpoint particular deficiencies of the speech processor. However, the total score is obtained by averaging the six main diagnostic scores.

4. RESULTS AND INTERPRETATIONS

The gross scores of the six critical phonemic features considered in the DRT are plotted in Fig 1. Score differences over subcategories are highlighted in the discussion, while subdivision of the scores according to the voicing state or specific phoneme cues are shown in Table 3.

Noteworthy are the consistent depressions on the voicing, graveness and sustention components for the conditions No. 5 and 6. In fact, these three features separate the RELP coders from the LPC vocoder.

The voicing feature distinguishes the voiced consonants from their unvoiced counterparts: /b/ from /p/, /d/ from /t/, /v/ from /f/, etc. For the vocoder conditions, there is a small but consistent bias towards the voicing absent state (i.e., voiced is more frequently perceived as unvoiced). This is due to a significant bias towards the friction absent state.

The graveness feature distinguishes /p/ from /t/, /b/ from /d/, /w/ from /r/, /m/ from /n/, etc. The graveness scores are the smallest for almost all conditions, and this wide gap is primarily due to the inherent difficulty in distinguishing the unvoiced consonant pairs /f/-/θ/ and /p/-/t/. The graveness scores exhibit a bias favoring the absent state, especially for the unvoiced and nonplosive sub-categories (see Table 3). This bias is by far larger for female than for male speakers.

We know that degradations on higher frequency components of voice signals affect the graveness and sibilation features most. Therefore, we can conclude that low scores on these features are also due to the inadequacy of the excitation signal fed into the synthesiser. This is more evident in the LPC algorithm, where the excitation signal is modeled in a rigid and poor way. Also in the RELP, however, the frequency components greater than 1 kHz in the excitation are regenerated in a synthetic manner using the baseband, and this approach is not efficient for certain phonemes and for speakers with high frequency energy concentration. This impairment may be mitigated by a better representation of the true full-band residual signal. Indeed, recent algorithms, such as Multi-pulse [9] or Regular-pulse LPC [10] and Vector Excited Coders [11,12,13], aim at improving the subjective quality at low bit-rates by exploiting a perceptually efficient excitation coding method.

The most significant difference between the three pairs of conditions (1-2, 3-4 and 5-6), is given by the sustention feature, which distinguishes the abrupt weak consonants from their sustained counterparts (/p/ from /f/, /b/ from /v/, /t/ from /θ/). The largest drop is observed from RELP to LPC conditions, and in fact sustention suffers the greatest impairment in vocoded speech. We note that for conditions 2, 3 and 4, the unvoiced sustention feature is affected by a bias towards the absent state, i.e., unvoiced sustained consonants become more like stops. For vocoder conditions 5 and 6, there is also a strong bias towards the voiced present state. This bias is primarily a result of pitch and gain coding, which made most voiced stop consonants (e.g. /b/) sound like continuants (e.g. /v/). Improvements on this effect can be obtained with faster frame update for unvoiced speech and better gain quantization.

The nasality feature, which distinguishes /n/ from /d/ and /m/ from /b/, is the best perceived feature for all conditions.

Scores for the feature compactness relate to the compact-diffuse attribute that serves to distinguish /y/ from /w/, /g/ from /d/, /k/ from /t/, /ʃ/ from /s/, etc. There are no significant differences between the voiced and unvoiced states of the compactness feature.

The sibilation feature, which distinguishes /s/ from /θ/, /ʃ/ from /k/, etc., shows a bias towards the absent state, indicating that strident consonants can be reproduced with mellow cues. This effect is due to deficiencies of the excitation signal, as discussed for the graveness feature.

The maximum degradation in going from the channel No.1 to the channel No.2 is about 5 points for the sustention feature. In particular, comparing the performance of RELP coders, we note that the error protection implemented on the RELP coder at 9.6 kbit/s seems to be more useful to preserve this feature along with sibilation (for the channel No.1) and graveness (for the channel No.2). In fact, large amounts of consonant feature information are carried in the duration and spectral characteristics of adjacent vowels, as well as in the acoustical manifestations of the consonants. Therefore, the error protection of spectral parameters from k(1) to k(4), particularly adequate for vowels, gives benefits also to certain consonants. Of course, loss of information in the upper frequency formants may cause significant degradations. The robustness of nasality for all the conditions, and of voicing for RELP configurations, is clearly evident. Also compactness, which depends on, among other things, the higher second-formant frequencies, appears somewhat robust for all the conditions.

Overall, the DRT scores show the remarkable robustness of the 9.6 kbit/s RELP system, even in case of multipath fading degradation.

The performance of the LPC system is mainly impaired on the voicing, graveness and sustention features, which are generally quite fragile in all vocoders and sensitive to various forms of speech degradation.

5. CONCLUSIONS

We have simulated in real time two speech coding systems at low bit-rates, suitable for mobile satellite communications. We have evaluated their robustness against typical channel degradations using the DRT facility, and got useful information to trade-off between important issues such as power, bandwidth, quality, complexity and delay. It turns out that a 9.6 kbit/s RELP coder is capable of ensuring very good intelligibility, provided that the most important parameters of the side-information be protected with a combination of bit interleaving and error-correcting codes. Short codes must be used. In fact, in addition to being simpler to decode, short codes are more adequate than long ones when the error probability of the channel is large. In particular, a (15,5) BCH code and a (12,4) code have proven to be suitable for our purposes.

Comparing the DRT scores, it results that two subjective categories are gained by the 9.6 kbit/s RELP over the 2.4 kbit/s LPC system. Indeed, the ability to yield fair quality at 2.4 kbit/s using conventional vocoders remains to be seen. Should this happen, however, it could allow an additional reduction of 4 in power and bandwidth.

Recent speech compression algorithms [9-13] provide high quality speech somewhere between 4 and 8 kbit/s, under ideal transmission conditions. Therefore, future problems to be addressed are those associated with their subjective performance in presence of environmental noise, channel errors and multipath fading.

ACKNOWLEDGMENTS

This study has been supported in part by ESA under ESTEC Contract No. 6098/84/NL. The authors would like to thank W. Kriedte, Agency's representative, for his co-operation. Thanks also go to E. Biglieri and G. Albertengo, from Politecnico of Torino, for the simulation of channel models and error-correcting codes.

REFERENCES

[1] M.Copperi et al., "Medium-rate speech coding simulator for mobile satellite systems", Final Report, ESTEC/Contract No.6098/84/NL, Jan.1986

[2] N.Dal Degan and V.Di Lago, "Design and test of a real-time floating point LPC vocoder", Proc. ICASSP, pp. 97-100, Apr. 1983

[3] W.D.Voiers, "Evaluating processed speech using the Diagnostic Rhyme Test", Speech Technology, pp. 30-39, Jan./Feb. 1983

[4] A.Huggins and R.Nickerson, "Speech quality evaluation using 'phoneme specific' sentences", J.A.S.A. Vol.77, pp.1896-1906, May 1985

[5] F.Yato et al., "Performance evaluation of voice coding schemes applicable to INMARSAT standard-B system", IEE 3rd Int. Conf. on Satellite System for Mob. Commun. Navigation, pp. 162-166, June 1983, London

[6] M.McLaughlin, D.Linder and S. Carney, "Design and test of spectrally efficient land mobile communications systems using LPC speech", IEEE Journ. Selected Areas Commun., Vol. 2, pp. 611-620, July 1984

[7] R.H.McCullough, "The binary regenerative channel", B.S.T.J., Vol. 47, pp. 1713-1735, 1968

[8] R.Viswanathan, A.Higgins and W.Russel, "Design of a robust baseband LPC coder for speech transmission over 9.6 kbit/s noisy channels", IEEE Trans. on Commun., Vol. 30, pp. 663-673, Apr. 1982

[9] B.Atal and J.Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates", Proc. ICASSP, pp. 614-617, May 1982, Paris

[10] P. Kroon, E. Deprettere and R. Sluyter, "Regular-pulse excitation: a novel approach to effective and efficient multipulse coding of speech", IEEE Trans. ASSP, Vol. 34, pp.1054-1063, Oct. 1986

[11] M.Copperi and D.Sereno, "Vector quantization and perceptual criteria for low-rate coding of speech", Proc. ICASSP, pp.252-255, Mar. 1985, Tampa (FL)

[12] M.Schroeder and B.Atal, "Code-Excited Linear Prediction (CELP): high-quality speech at very low bit rates", Proc. ICASSP, pp. 937-940, Mar. 1985, Tampa (FL)

[13] M.Copperi and D.Sereno, "CELP coding for high quality speech at 8 kbit/s", Proc. ICASSP, pp. 1685-1688, Apr. 1986, Tokyo

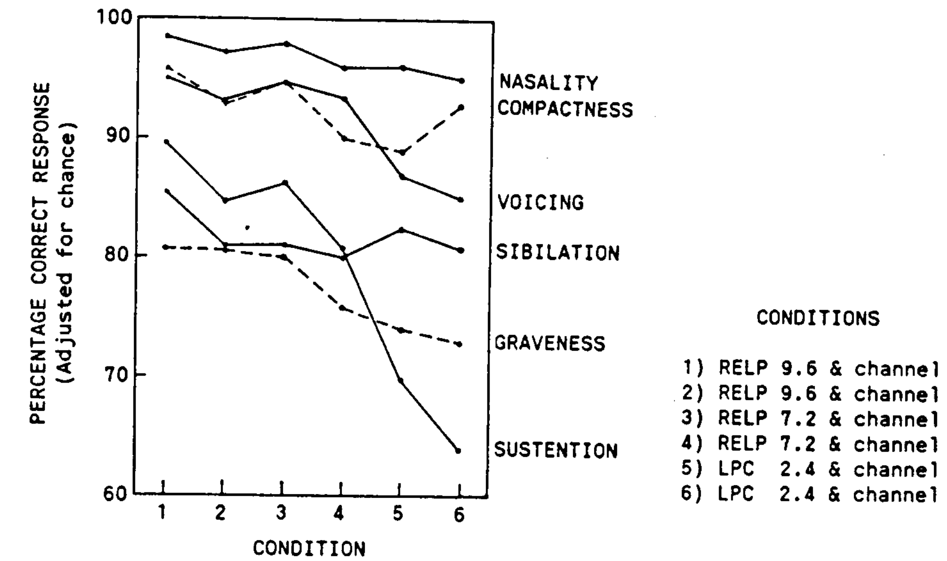


Fig.1 - DRT scores

FEATURES	PHONEMES																						
	m	n	v	ð	z	ʒ	$\hat{ʒ}$	b	d	g	w	r	l	j	f	θ	s	ʃ	$\hat{ʃ}$	p	t	k	h
Voicing	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
Nasality	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Sustention	-	-	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	+
Sibilation	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-
Graveness	+	-	+	-	-	0	0	+	-	0	+	-	0	0	+	-	-	0	0	+	-	0	0
Compactness	-	-	-	-	-	+	+	-	-	+	-	-	0	+	-	-	-	+	+	-	-	+	+

Tab. 1 - Consonant taxonomy used in DRT [3]

VOICING voiced-unvoiced	NASALITY nasal-oral	SUSTENTION sustained-interrupted
veal-feel	meat-bear	vee-bee
bean-peen	need-deed	sheet-cheat
gin-chin	nip-dip	vill-bill
dint-tint	moot-boot	thick-tick
zoo-Sue	news-dues	foo-pooh
dune-tune	moan-bone	shoes-choose
goat-coat	neck-deck	those-doze
dense-tense	mad-bad	shaw-chaw
jock-chock	knock-dock	fence-pence
SIBILATION sibil.-unsibil.	**GRAVENESS grave-acute**	**COMPACTNESS compact-diffuse**
cheep-keep	weed-reed	yield-wield
jilt-gilt	peak-teak	key-tea
sing-thing	bid-did	hit-fit
chew-coo	fin-thin	you-rue
juice-goose	moon-noon	ghost-boast
sole-thole	pool-tool	coop-poop
chair-care	fore-thor	yawl-wall
jab-dab	bond-dong	got-dot
zee-thee	wad-rod	shag-sag

Tab. 2 - Sample of DRT stimulus words [3]

FEATURES	CONDITIONS 1	2	3	4	5	6
VOICING	95.1	93.0	94.7	93.4	86.9	85.0
frictional	91.8	89.1	91.0	89.1	82.0	79.7
nonfrictional	98.4	96.9	98.4	97.7	91.8	90.2
NASALITY	98.4	97.3	98.0	95.9	96.1	95.1
grave	98.8	96.9	98.0	94.5	95.7	97.7
acute	98.0	97.7	98.0	97.3	96.5	92.6
SUSTENTION	89.5	84.4	86.3	80.5	69.7	63.9
voiced	86.3	80.1	85.9	77.0	70.3	57.0
unvoiced	92.6	88.7	86.7	84.0	69.1	70.7
SIBILATION	85.5	81.1	81.3	80.1	82.4	80.7
voiced	91.4	88.3	86.7	83.6	85.9	87.9
unvoiced	79.7	73.8	75.8	76.6	78.9	73.4
GRAVENESS	80.5	80.5	80.1	75.8	74.0	72.9
voiced	94.5	96.1	93.0	87.1	87.1	83.2
unvoiced	66.4	64.8	67.2	64.5	60.9	62.5
plosive	84.0	86.3	84.4	80.9	74.2	76.2
nonplosive	77.0	74.6	75.8	70.7	73.8	69.5
COMPACTNESS	95.7	92.8	94.5	90.0	89.1	93.0
voiced	95.7	90.2	95.3	92.6	93.0	96.1
unvoiced	95.7	95.3	93.8	87.5	85.2	89.8
TOTAL SCORE	90.8	88.2	89.2	85.9	83.0	81.7
STD. ERROR	.86	.94	.65	1.03	.72	.77

Tab. 3 - DRT scores of main features and sub-categories

A RELATIONSHIP BETWEEN THE QUALITY OF VOCODED SPEECH AND ITS COMPRESSION RATIO

KASTYTIS RATKEVIČIUS ALGIMANTAS RUDŽIONIS

Speech Research Laboratory
Kaunas Polytechnical Institute
Kaunas, Lithuania, USSR 233028

ABSTRACT

A 24-channel vocoder was used to study the quality of vocoded speech under the effects of its compression variables - the number of spectral parameters n, sampling period T and bit number m in vocoder spectral parameters. Syllable intelligibility S and speaker recognition σ (identification) were used as measures of the quality. To reduce the number of spectral parameters the method of averaging over subsequent amplitude spectrum samples (AS) is suggested.

INTRODUCTION

Major variables of the compression ratio in a channel vocoder are: the number of spectral parameters n (the number of channel signals, which represents the envelope of the short-time spectrum of the speech signals), sampling period T (the sampling interval of any spectral parameter), bit number m (the number of quantization bits per one spectral parameter). Design of vocoders with pre-given properties demands a proper knowledge of relations between the quality of vocoded speech and the above variables. We are not aware of any efficient and reliable method of evaluating vocoded speech quality. Most reseachers rely on intelligibility and speaker recognition in their judgements over processed speech. There is a number of publications on intelligibility, but none of them reflects properly relations between vocoded speech intelligibility and variables n, m, T. As to speaker recognition, we can mention just one study dealing with the evaluation of recognition accuracy of LPC speech /1/. Tape recordings of 24 speakers conversing over an unprocessed channel and over an LPC voice processing system with the rate 2400 bit/s were subjected to listening tests. The listeners were 24 co-workers who attempted to identify each speaker from a group of about 40 people working in the same branch. The average duration of the speech samples was 29,8 s. Recognition accuracy was 88% for unprocessed speech, and 69% for LPC speech. No evaluation of the effect of compression ratio on the speaker recognition accuracy was made.
Note, that most of industrial vocoders use differential pulse code modulation (DPCM), in which the reference parametric signal is coded with a 3-bit logarithmic code, other signals-by 2-bit DPCM /2, 3/. This type of coding is very popular, yet, the relation between intelligibility of vocoded speech, as its main quality measure, and the number of quantization bits per one spectral parameter is interesting from the point of view of the relative information of variable m in comparison with n and T.

We also attempted to find a simple and reliable method of reducing the number of spectral parameters. This may be done by averaging subsequent samples of the amplitude spectrum (AS).

ACCURACY OF SPECTRUM REPRODUCTION

To evaluate the efficiency of the suggested averaging approach, we performed a comparative analysis of spectrum reproduction by the two methods of reducing the number of spectral parameters. A samples of amplitude spectrum of the vocoder analyser output signal were subjected to harmonic approximation (approximate representation of the spectral envelope by means of a Fourier series) by the first method, and a certain number of subsequent components was averaged by the second method, that is several subsequent samples of the amplitude spectrum from the vocoder synthesizer output signal were replaced by the average value of the analyser output samples. The test was performed on a micro-computer-aided 24-channel vocoder with a high syllable intelligibility of the vocoded speech (average score 94,3% at data rate 4800 bit/s). Its freguency range was from 100 Hz to 8 kHz. The analyser was equiped with 6-th order Bessel band-pass filters having 3 dB attenuation at 25 Hz. In the synthesizer 20 narrow-band 2-nd order filters with outputs combined in an antiphase summation were used to cover the speech band to 5 kHz and 4 wideband filters were used in the upper freguency range. Modulation was done by digital-analogous converters.From identical speech samples, the absolute average error of one spectral component-error of spectrum reproduction - was found by

$$\delta = \frac{1}{K}\frac{1}{n_{max}}\sum_{k=1}^{K}\sum_{i=1}^{n_{max}}|F_{Ak}(i)-F_{Sk}(i)|$$

where $F_{Ak}(i)$, $F_{Sk}(i)$ - the i-th sample of the k-th spectrum frame on the analyser output, and on the synthesizer input, respectively. Fig.1 presents the dependence of δ on the number of spectral parameters n for the two methods of reducing this number.

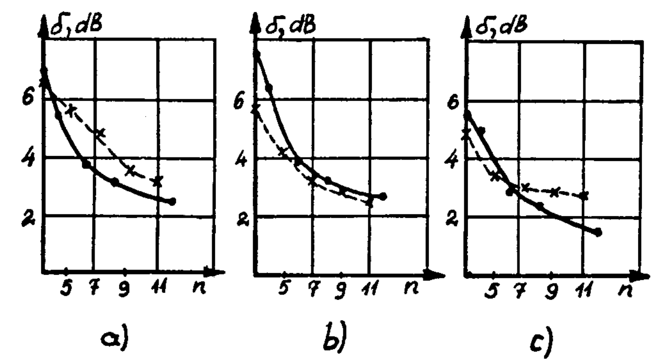


Fig.1. Accuracy of spectrum reproduction for long utterance (a), long vowels /a,i/ (b) and fricative consonant /ʃ/ (c).
• - averaging
× - harmonic approximation

Approximation by harmonic functions is only preferable for vowel phonemes, and long utterances (sentences) are more accurately reproduced by averaging subsequent spectrum samples. Human perception tests on the two methods suggest their comparative effects of spectrum reproduction. Further we restrict ourselves by the more simple method of averaging.

INTELLIGIBILITY OF VOCODED SPEECH

To evaluate the quality of vocoded speech, its syllable intelligibility S was evaluated as a more objective factor, as compared to intelligibility of phrases and words. For each test, five tables of phonetically balanced syllables of Russian words (total 250 syllables) were recorded by one male reader and processed in the vocoder. Samples of spectrum cutoffs on the output of the analyzer were microcomputer-processed to reduce the number of quantization bits per one spectral parameter and the number of spectral parameters. The necessary sampling period was controlled manually by the switch. Samples of compressed spectrum cut-offs were fed to the synthesizer, as samples of the following spectrum cut-offs from the output of the analyser were fed to the computer. Processed syllables were recorded on a magnetic tape and played before three listeners. Syllable intelligibility for separate listeners and average intelligibility of vocoded speech were then determined. Two more simple ways of reducing bit number m were tested: a) transfer of amplitude spectrum samples (TS), when several subsequent amplitude spectrum samples on the synthesizer input $F_S(i)$ are replaced by a single amplitude spectrum sample on the analyser output, $F_A(i)$ and b) deletion of samples (DS), when separate values F_S=0. The determined relations between syllable intelligibility S and the number of spectral parameters n for TS, DS and AS methods are shown in Fig.2.

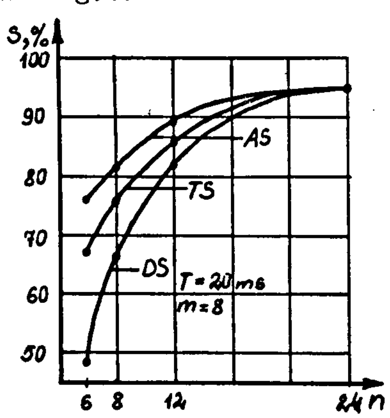


Fig.2. Dependence of syllable intelligibility S on the number of spectral parameters n

Deviation from the natural speech was observed at n<I2. The undoubtful advantage of the averaging method was stated and further it was used in the evaluations of intelligibility.

A relations between syllable intelligibility S of vocoded speech and the number of quantization bits per one spectral parameter m for given numbers of spectral parameters n are shown in Fig.3a.A distinct deviation from natural speech occurs at m< 3.

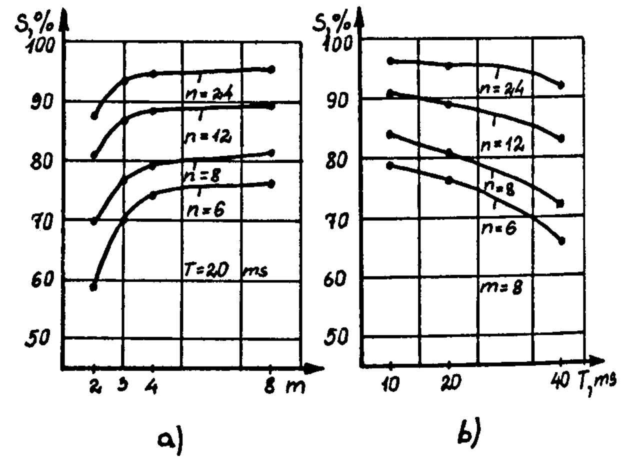


Fig.3. Dependences of syllable intelligibility S on variables m(a), T(b).

For the same number of spectral parameters n, relations between syllable intelligibility S and sampling period T were measured (Fig.3b). Deviation from natural speech at T=40 s is mainly due to the failure in the synthesis of short sounds.

SPEAKER RECOGNITION

The evaluation of speaker recognition Ɉ comes from two tests. First an attempt was made to find a relation between speaker recognition from unprocessed speech and vocoder-processed speech and the duration of speech sample. Speech samples were collected from 11 known speakers and 4 unknown speakers. 5 monosyllabic words each of 0.5 s average duration, 5 polysyllabic words - 1.5 s and 5 phrases - 4 s were used. Each sequence of samples was chosen at random, recorded on a magnetic tape and played before 11 listeners. A warning concerning the unknown speakers was made. The listeners were carefully instructed not to chek off names or to use any process of elimination because some speakers were sampled more than once. Unprocessed speech, vocoded speech and monotonous vocoded speech were tested.The values of compression variables were: n==24, m=8, T=40 ms. The relations of speaker recognition Ɉ and the duration of speech sample t are shown in Fig.4a.

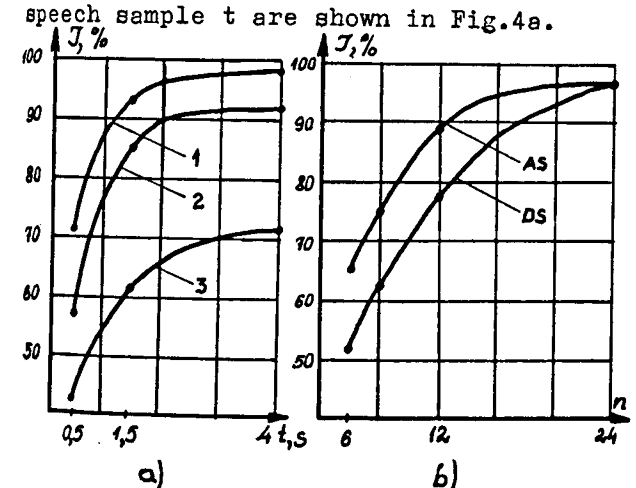


Fig.4. Speaker recognition Ɉ in relation to the duration of speech sample t (a) (1-unprocessed, 2-vocoded, 3- monotonous vocoded speech) and to the number of spectral parameters n (b).

The absence of fundamental frequency of the voice results in a 20% decrease in the recognition accuracy. For a successful identification and verification of a speaker from vocoded speech, just a phrase or a polysyllabic word may be used without appreciable loss of accuracy.

The second test was done by 11 known speakers. Only phrases were read, and only vocoded speech with different numbers of spectral parameters n, was evaluated. The resultant relations of speaker recognition Ɉ and the number of spectral parameters n for two methods of reducing this number are shown in Fig.4b. The recognition accuracy becomes significantly lower at n<12. The comparison supports the advantages of the averaging method.

CONCLUSIONS

The presented relations of vocoded speech syllable intelligibility S to the number of spectral parameters n, sampling period T and the number of quantization bits per one spectral parameter m, as well as the relations of speaker recognition Ɉ to the number of spectral parameters n open the ways towards designing vocoders for pregiven properties of the processed speech, which can be achieved by choosing proper design parameters. Test evaluations of the quality of vocoded speech may be described by the limiting values of variables n, m, T, which are: n≥12, m≥3, T≤40 ms. Vocoded speech becomes significantly unnatural whenever one of the variables exceeds it limiting value. We underline the advantages of the method of reducing the number of spectral parameters by averaging subsequent samples of speech amplitude spectrum. Problems of speaker recognition and verification from vocoded speech may be solved on single utterunce from 2 to 4 s.

REFERENCES

/1/ A. Schmidt-Nielsen, K. Stern, "Identification of known voices as a function of familiarity and narrow-band coding", JASA, Vol.77, pp.658-663, 1985.

/2/ B. Gold, P.E. Blankenship, R.J. McAulay," New applications of channel vocoders", IEEE Trans. ASSP, Vol.29, No.1, pp.13-32, 1981.

/3/ J.N. Holmes, "The JSRU channel vocoder", IEE Proc. Communications, Radar and Signal Processing, 127, Pt.F, pp.53-60, 1980.

КОДИРОВАНИЕ РЕЧЕВЫХ СИГНАЛОВ ДЛЯ ЦЕЛЕЙ ЭЛЕКТРОДНОГО ПРОТЕЗИРОВАНИЯ СЛУХА

Л.В. ЛЕСОГОР

Институт физиологии им.И.П.Павлова АН СССР, Ленинград.

РЕЗЮМЕ

Описаны психоакустические эксперименты с восприятием кодированных речевых сигналов нормально слышащими испытуемыми. Кодирование осуществлялось таким образом, что исходный сигнал преобразовывался в последовательность коротких биполярных импульсов, модулированных по амплитуде огибающей речевого сигнала.

Исследовалось влияние различных параметров кодирования на словесную разборчивость.

ВВЕДЕНИЕ

Электродное протезирование слуха или кохлеарная имплантация является одним из новейших направлений реабилитации слуха при полной глухоте. Основная задача, которая стоит перед исследователями, занимающимися электродным протезированием состоит в том, чтобы осуществить такое преобразование речевого сигнала в электрические стимулы, которое с одной стороны сохраняло бы как можно больше информации о речевом сигнале, с другой – удовлетворяло тем условиям, которые возникают в связи со спецификой электрического возбуждения слухового нерва.

Все имеющиеся в настоящее время системы электродного протезирования разделяют на одноканальные и многоканальные. Принципиальное отличие многоканальной системы от одноканальной состоит в том, что многоканальные системы протезирования, имеют несколько параллельных каналов, разделяющих акустический сигнал на спектральные полосы посредством фильтров, и таким образом могут обеспечить изображение речевых сигналов, более близкое к естественному. При одноканальном же стимулировании необходимо произвести такое кодирование акустического сигнала, которое позволит выделить из него признаки, содержащиеся во временной картине [I]. Одноканальные системы обладают рядом существенных достоинств, и именно поэтому, в настоящее время они получили наибольшее распространение. Основные преимущества заключаются в следующем: I. Наличие одного канала и, следовательно, одного электрода, значительно упрощает конструкцию и, что особенно важно, имплантацию кохлеарных протезов. 2. Только одноканальная стимуляция имеет смысл в тех случаях, когда у больного остается небольшое количество сохранившихся (непораженных) волокон слухового нерва. 3. Временная структура стимулирующего сигнала является гораздо информативнее, чем это предполагалось до последнего времени; с её помощью может передаваться не только просодическая информация, но и информация о фонемах, а также можно достичь удовлетворительного распознавания как отдельных согласных и односложных слов, так и слитной речи [4]. Более эффективными являются такие одноканальные системы электродного протезирования, в которых преобразование осуществляется таким образом, что информация о частотном составе речевого сигнала, кодируется величиной межимпульсных интервалов, а информация об интенсивности – либо длительностью, либо амплитудой импульсов [5].

В настоящее время разработано и испытывается много типов речевых преобразователей, проводятся их сравнительные исследования, однако до сих пор еще нет четких данных, касающихся выбора наиболее оптимальных параметров кодирования: частоты среза фильтра, длительности биполярных импульсов и т.д. Это можно объяснить тем, что исследования, проводимые на пациентах с имплантируемыми электродами усложняются целым рядом практических трудностей: а) число возможных пациентов мало (I или 2); б) имплантируемое устройство налагает ограничения на сигнал стимуляции; в) результаты зависят от таких неконтролируемых факторов, связанных с патологией улитки, как число и характеристическая частота уцелевших слуховых нейронов. Поэтому нами была предпринята попытка решить эту задачу с помощью испытуемых, обладающих нормальным слухом, исходя из того, что оптимальность параметров выбранная испытуемыми с нормальным слухом будет также действительна и для глухих пациентов. Известно, что использование нормальнослышащих испытуемых для сравнительного анализа схем речевого кодирования в работе [2] дало хорошие результаты.

ОПИСАНИЕ КОДИРУЮЩЕГО УСТРОЙСТВА И ПРОЦЕДУРА ИЗМЕРЕНИЙ

На основании вышеизложенного нами было разработано устройство, позволяющее кодировать как речевой, так и любой акустический или электрический сигнал одним, как нам кажется, из наиболее оптимальных способов, при котором исходный сигнал преобразуется в последовательность коротких биполярных импульсов, модулированных по амплитуде огибающей речевого сигнала.

Преобразователь выполнен в виде отдельного устройства, на вход которого сигналы могут подаваться либо с микрофона, либо от любого источника электрических сигналов. В устройстве имеется фильтр пропускания низких частот (ФНЧ), частоты среза которого устанавливаются переключателем и принимают значения от 0.5 кГц до 3.0 кГц ступенями через 0.5 кГц. Динамический диапазон амплитуды выходных импульсов определяется амплитудной характеристикой входящего в него логарифмического усилителя, который обладает компрессией I:I0. Амплитудная огибающая переменного сигнала выделяется детектором и затем сглаживается с помощью интегратора. Формирование импульсов осуществляется посредством триггера Шмитта и одностабильных мультивибраторов. На выходе, таким образом, получаются амплитудно-модулированные биполярные импульсы заданной длительности, моменты появления которых соответствуют моментам пересечения нуля (в положительной фазе) сигнала с выхода ФНЧ.

Процедура измерения состояла в следующем: записанные на грампластинку слова с речевого аудиометра АР-03 поступали на кодирующее устройство, телефоны. Прослушивание производилось в звукоизолированной камере, моноурально, испытуемыми с нормальным слухом. Использовались специально составленные таблицы слов Г.И.Гринберга, Л.Р.Зиндера и Л.В.Нейман. Были проведены 4 серии экспериментов по определению наиболее оптимальных значений следующих параметров электрических стимулов: частоты среза ФНЧ, f_c; длительности биполярных импульсов, $t_и$; постоянной времени интегрирования, τ; интенсивности подаваемых стимулов, L : С этой целью были сняты зависимости разборчивости речевых тест-таблиц от значения этих параметров. Разборчивость определялась по количеству правильно воспринятых и записанных испытуемыми слов (из 60 предъявленных) и выражалась в процентах.

$$p = \frac{1}{m \cdot n} \sum_{\substack{i=1 \\ k=1}}^{\substack{k=n \\ i=m}} p_{ik} [\%]$$

где P_{ik} – процент правильно принятых слов i – ым аудитором в k – ой таблице, m – число испытуемых, n – число прослушанных таблиц.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Данные, полученные в экспериментах по определению разборчивости речевого теста в зависимости от частоты среза фильтра нижних частот, (при постоянной времени сглаживания 5 мсек и длительности биполярных импульсов I00 и 500 мксек), приведены на рис. I.

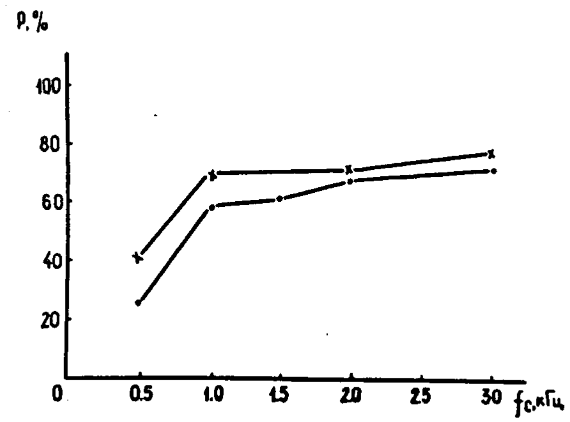


Рис. I. Зависимость словесной разборчивости от частоты среза ФНЧ.

По оси ординат – разборчивость в процентах; по оси абсцисс – частота в кГц (усредненные данные для 5 испытуемых).
х – t_n = I00 мксек; 0 – $t_и$ = 500 мксек τ = 5 мсек.

Можно видеть, что разборчивость резко возрастает с расширением полосы пропускания сигнала до I.0 кГц; при дальнейшем увеличении частоты среза разборчивость меняется незначительно. Исследования нейронных ответов при электрической стимуляции также указывают на ограниченный частотный диапазон, не превышающий I кГц, при котором сохраняется информация о периоде колебаний стимулов. Это позволяет предположить, что если нормальнослышащий испытуемый способен различать речевые стимулы, прошедшие НЧ фильтрацию с частотой среза I.0 – I.5 кГц, то и пациент с одноканальной стимуляцией внутреннего уха, диапазон воспринимаемых частот которого ограничен такой же частотной областью сможет распознавать речевые стимулы, используя такие параметры, как частоту основного тона и частоту I-й и 2-й формант, лежащих в пределах этой области.

Во второй серии экспериментов исследовалось влияние постоянной времени сглаживания амплитудной огибающей, которая принимала значения, равные 5, 20 и 50 мсек, на разборчивость речевого теста при постоянных значениях частоты среза (I.5 кГц) и длительности импульсов (500 мксек). Результаты эксперимента показали, что постоянная времени сглаживания очень незначительно влияет на разборчивость. Анализ ошибок ответов указывает на то, что увеличение постоянной времени сглаживания влияет на восприятие лишь начальной фонемы в слове. Поэтому при τ = 50 мсек увеличивается число ошибок в распознании начальной фонемы. При значениях τ =5мсек качество звучания ухудшается за счет восприятия щелчка.

На основании вышесказанного можно сделать вывод о том, что наиболее оптимальными с точки зрения правильного распознавания, являются значения постоянной времени сглаживания I0–20 мсек.

В третьей серии экспериментов определялась зависимость словесной разборчивости от длительности биполярных импульсов. Длительность импульсов менялась в пределах от I00 мксек до I000 мксек. Данные были получены для трех испытуемых. Исходя из результатов предыдущих серий экспериментов частота среза ФНЧ была выбрана равной I.5 кГц, постоянная времени сглаживания амплитудной огибающей исходного сигнала I0 мс. Из эксперимента следует, что с увеличение длительности биполярных импульсов до 500 мксек, процент правильного восприятия монотонно возрастает, а затем остается постоянным, равным ≈ 70%. Наши результаты хорошо согласуются с данными, полученными у пациентов с имплантированными электродами [3].

Из вышесказанного следует, что оптимальная длительность биполярных импульсов составляет около 500 мксек.

В четвертой серии экспериментов измерялась зависимость словесной разборчивости от уровня подаваемых речевых стимулов. Т.е. определялось, существует ли какой-то наиболее оптимальный уровень интенсивности речевых стимулов, при котором разборчивость получается наибольшей. Уровень звукового давления измерялся в дБ над порогом слышимости отдельно для каждого испытуемого. Как можно видеть из рисунка 2 даже при низком уровне интенсивности стимулов (I0 дБ) разборчивость достаточно высока; для 2-х испытуемых она превышает 50% и намного выше начального уровня раборчивости (20%) для третьего испытуемого.

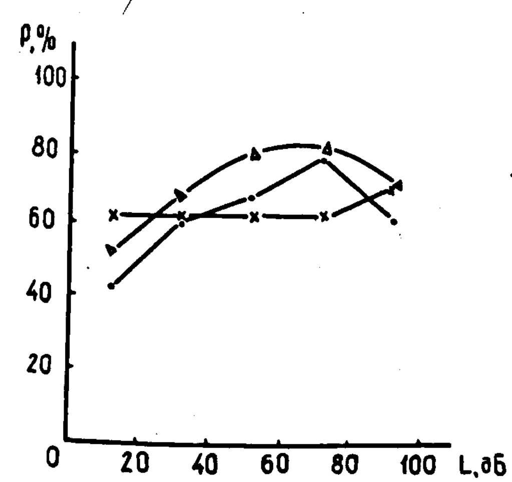


Рис. 2. Зависимость словесной разборчивости от уровня подаваемых речевых стимулов для трех испытуемых.

Δ – исп. А; х – исп. Б; • – исп. В при f_c = I.5 кГц и τ = I0 мсек.

Наиболее оптимальным уровнем, при котором разборчивость достигает 82% (исп. А и Б), оказался уровень 70 дБ. При дальнейшем увеличении интенсивности сигнала происходит некоторое ухудшение разборчивости, вызванное слишком высоким уровнем интенсивности стимулов, близким к болевому порогу. Таким образом, результаты данной работы позволили посредством психоакустических экспериментов произвести предварительный выбор оптимальных параметров электрических сигналов, что является необходимым шагом на данном этапе исследований в области электродного протезирования, т.к. позволит более обоснованно подходить к созданию аппаратуры для кодировани речи и сократить время постоперационного тестирования больных с имплантируемыми электродами.

ЛИТЕРАТУРА

I. Люблинская В.В. Слуховое восприятие у человека при электрическом раздражении слуховой системы. Раздражение периферических отделом слуховой системы. В кн.: Электродное протезирование слуха. Л.,: "Наука", I984, стр. I47–I7I.
2. Blamey P.Y., Martin L.F.A., Clark G.M. A comparison of three speech coding strategies using an acoustic model of a cochlear implant. – J.Acoust. Soc. Amer., 1985, v.77 (1), p. 209-217.
3. Burian K., Hochmair E., Hochmair-Desoyer I.J., Lessel M.B. Designing of the experience with multichannel cochlear implants. – Acta otolaryngol., 1979, v.86, p. 190-195.
4. Hochmair-Decoyer I.J., Hochmair E.S., Fischer R.E., Burian K. Cochlear prosthesis in use: Recent speech comprehension results. – Arch.Oto-Rhino-Laryngol., 1980, v.229, p. 81-98.

МЕТОД ПОВЫШЕНИЯ КАЧЕСТВА ЗВУЧАНИЯ СИНТЕЗИРОВАННОГО РЕЧЕВОГО СИГНАЛА В ЦИФРОВОМ ВОКАДЕРЕ С ПРЕДСКАЗАНИЕМ

С.Ф.Лихачев, М.В.Назаров, Ю.Н.Прохоров

Московский электротехнический
институт связи

Р е ф е р а т

Предложен новый метод оценивания параметров речевых сигналов на основе метода обновляющего процесса (ОП). По сравнению с традиционным методом наименьших квадратов (МНК), разработанный метод позволяет повысить качество звучания синтезированного речевого сигнала в цифровом вокодере, работающем при действии шумов умеренной интенсивности.

I. Введение

Предметом настоящего доклада является исследование и разработка методов повышения эффективности рекуррентных алгоритмов оценивания параметров, обеспечивающих повышение качества цифровой передачи речи при низких скоростях 2400...4800 бит/с и наличии входных шумов умеренной интенсивности.

В настоящее время наиболее перспективными в этом направлении являются системы передачи с предсказанием [I,2,3].

Известно, что в вокодерах с линейным предсказанием не учитываются особенности слухового восприятия в частотной области.

В работах [I,2,4,5] и других доказано, что область низких частот наиболее важна для слуха. В тоже время методы линейного предсказания, построенные в рамках метода наименьших квадратов, обеспечивают малую ошибку в описании высокочастотной области спектра, к которой ухо менее чувствительно. Исходя из этого, отличие спектров исходного и синтезированного речевых сигналов оказывается довольно значительным. Низкое качество восстановленного сигнала в низкоскоростных системах с линейным предсказанием обусловлено тем, что погрешность предсказания содержит информацию о спектре, которую не может извлечь анализатор системы передачи, построенный на методе наименьших квадратов. Поэтому большой научный и технический интерес представляет разработка метода такого изменения спектрального состава восстановленного речевого сигнала, при котором учитываются особенности слухового восприятия.

2. Метод оценивания параметров предсказания.

Введем авторегрессионную модель сигнала

$$x_t=\vec{a}^T\vec{\varphi}(x_{t-1})+\xi_t, \qquad (I)$$

где $\vec{a}^T=(a^{(1)},\dots,a^{(p)})$ – вектор параметров авторегрессии; $\vec{\varphi}(x_{t-1})$ – функция регрессии; ξ_t – порождающий процесс; p – размерность модели.

Если сигнал и модель стохастически

эквивалентны, то обновляющий процесс ν_t оказывается последовательностью некоррелированных случайных величин, обладающих теми же характеристиками, что и процесс ξ_t. Поэтому синтез сигнала на приемной стороне системы передачи можно представить как прохождение обновляющего процесса ν_t через линейное звено с передаточной функцией $K(\omega)$. Отклонение формантных максимумов в спектре речевого сигнала заметно на слух, если оно превышает +IдБ. Таким образом, из физических соображений следует, что показатель качества, отражающий спектральные свойства синтезированного аналогового речевого сигнала, должен представлять собой меру отклонения спектров исходного ($G_{xx}(\omega)$) и синтезированного ($\hat{G}_{xx}(\omega)$) речевых сигналов. Такую меру можно представить в виде:

$$\rho=\int\left|[G_{xx}(\omega)-K(\omega)\cdot G_{\nu\nu}(\omega)]\cdot M(\omega)\right|d\omega, \qquad (2)$$

где $M(\omega)$ – функция веса.

Функция потерь (2) не позволяет получить аналитически простые алгоритмы оценивания параметров речевого сигнала. Пусть $M=0$, если $G_{xx}(\omega)-K(\omega)\cdot G_{\nu\nu}(\omega)<0$. Тогда можно записать

$$\rho=\left|\int[G_{xx}(\omega)-K(\omega)\cdot G_{\nu\nu}(\omega)]\,M(\omega)\cdot d\omega\right|. \qquad (3)$$

Используя преобразование Винера-Хинчина меру ρ (3) теперь можно представить в виде:

$$\rho=\left|\int[B_{xx}(\tau_1)-q(\tau_1)B_{\nu\nu}(\tau_1)]\,\mu(\tau_1)\,d\tau_1\right| \qquad (4)$$

где $\mu(\tau_1)$ и $M(\omega)$ связаны преобразованием Фурье; $q(\tau_1)=\frac{\ell(\tau_1)}{\mu(\tau_1)}$ и $\ell(\tau_1)$ – преобразование Фурье от $K(\omega)\cdot M(\omega)$.

Заменяя интеграл интегральной суммой и подставляя в нее вместо $B_{xx}(\tau)$ ее оценку, получим показатель качества в дискретном времени:

$$\mathcal{J}_{\nu\nu}(\vec{a})=\left|\sum_{\tau=0}^{n}w(\tau)\sum_{t=1}^{N}[x_t-\vec{a}^T\vec{\varphi}(x_{t-1})][x_{t+\tau}-\vec{a}^T\vec{\varphi}(x_{t+\tau-1})]\right| \qquad (5)$$

где $w(\tau)$ – весовая последовательность.

Задача оценивания параметров может быть сформулирована следующим образом: по наблюдаемой последовательности x_t или $z_t=x_t+V_t$, $t=1,2,\dots N$ и априори заданной модели сигнала (I) определить наилучшую $\vec{m}_N$ из условия:

$$\vec{m}_N=\underset{\vec{a}}{\arg\min}\,\mathcal{J}_{\nu\nu}(\vec{a}), \qquad (6)$$

где V_t – шумовая последовательность; $\vec{m}_N$ – оценка вектора $\vec{a}$.

Метод отыскания оценок параметров модели авторегрессии минимизацией целевой функции $\mathcal{J}_{\nu\nu}(\vec{a})$ является развитием метода обновляющего процесса [I].

Из (5) можно получить оптимальную в смысле (6) оценку

$$\vec{m}_N=\left[\sum_{\tau=0}^{n}w(\tau)\sum_{t=1}^{N}\vec{\varphi}(x_{t-1})\cdot\vec{\varphi}^T(x_{t+\tau-1})\right]^{-1}\sum_{\tau=0}^{n}w(\tau)\sum_{t=1}^{N}x_{t+\tau}\,\vec{\varphi}(x_{t-1}) \qquad (7)$$

Используя лемму об обращении матриц получим рекуррентные выражения для оценок:

$$\vec{m}_t=\vec{m}_{t-1}+\vec{\gamma}_{t-1}\vec{\varphi}(x_{t-1})\sum_{\tau=0}^{n}w(\tau)[x_{t+\tau}-\vec{m}^T_{t-1}\vec{\varphi}(x_{t+\tau-1})];$$

$$\vec{\gamma}_t=\vec{\gamma}_{t-1}-\vec{\gamma}_{t-1}\vec{\varphi}(x_{t-1})\Big[1+\sum_{\tau=0}^{n}w(\tau)\cdot\vec{\varphi}(x_{t+\tau-1})\vec{\gamma}_{t-1}\vec{\varphi}(x_{t-1})\Big]^{-1}\cdot\sum_{\tau=0}^{n}w(\tau)\cdot\vec{\varphi}(x_{t+\tau-1})\cdot\vec{\gamma}_{t-1}, \qquad (8)$$

с начальными условиями

$$\vec{m}_0=E\vec{a};\ \gamma_0=\varepsilon^{-1}\vec{I},\ \varepsilon\to 0,\ \forall t\in N,\ \gamma_t\ (p\times p)$$

При стохастической эквивалентности сигнала и модели оценка (7),(8) совпадает при $N\to\infty$ с асимптотической оценкой МНК, но в отличии от нее в линейном случае оказывается несмещенной при стационарных шумах с равномерным спектром, так так, например, при p = I, $w(\tau)=\delta_{\tau\tau_0}$, имеем из (7):

$$\lim_{N\to\infty}m_N=\frac{B_{xx}(\tau_0+1)+B_{xx}(\tau_0+1)}{B_{xx}(\tau_0)+B_{VV}(\tau_0)}=\frac{B_{xx}(\tau_0+1)}{B_{xx}(\tau_0)}. \qquad (9)$$

Свойство асимптотической несмещенности сохраняется для линейной модели авторегрессии при любом τ_0, но может нарушаться в нелинейном случае.

Сложность технической реализации алгоритма (8) обусловлена необходимостью вычисления матрицы $\vec{\gamma}_t$. Для упрощения вычислений был предложен приближенный алгоритм оценивания:

$$\vec{m}_t = \vec{m}_{t-1} + \vec{\Gamma}_t \sum_{\tau=0}^{n} \omega(\tau)\, \vec{\varphi}(x_{t-1}) [x_{t+\tau} - \vec{\varphi}^T(x_{t+\tau-1}) \vec{m}_{t-1}], \quad (10)$$

где $\vec{\Gamma}_t = \Gamma_0 / t$ – матрица коэффициентов. Можно показать, что $\lim_{t\to\infty} \vec{m}_t = \vec{a}$.

3. Экспериментальное исследование.

Экспериментальная проверка алгоритма оценивания параметров проводилась для линейной и нелинейной моделей предсказания. В случае линейной модели показано, что разработанные алгоритмы по сравнению с алгоритмами МНК обеспечивают меньшее смещение допредельных оценок параметров в шумах с равномерным спектром. На рис.I показано смещение оценки $\vec{m}_t$ параметра a в шумах (при истинном значении a=-0,8) для линейной модели авторегрессии первого порядка, при различных отношениях сигнал-шум.

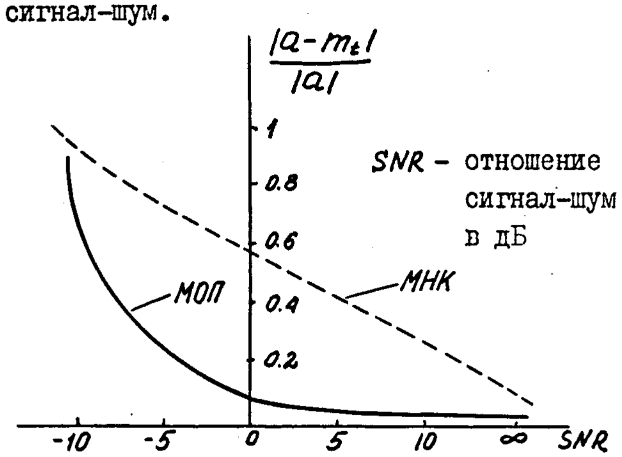


Рис.I Смещение оценок параметров.

Из рисунка видно, что при SNR= 5...10 дБ, смещение оценок в 2...3 раза меньше, чем при МНК.

В качестве нелинейной модели рассмотрена модель предсказания, в которой регрессии $\vec{\varphi}(x_{t-1})$ представлена в виде ряда по функциям Уолша:

$$\vec{\varphi}(x_{t-1}) = \sum_{k=1}^{m} \sum_{i=0}^{N-1} a_k^{(i)} \cdot wal^{(i)}(x_{t-1})$$

где $\{a_k^{(i)}\}, k=1,\dots,m;\ i=0,\dots,N-1$ – параметры нелинейного предсказания; $wal^{(i)}(x)$ – функция Уолша i -го порядка. Такое представление позволяет учесть негауссовское распределение вероятностей сигнала.

Алгоритмы оценивания параметров в случае нелинейной модели обеспечивают "обеление" погрешности предсказания. В частности, в ней подавляются импульсы основного тона. На рис.2 представлена корреляционная функция погрешности предсказания речевого сигнала получения по МНК МОП. Отрезком показан 95%-ый доверительный интервал; $T_{от}$ – период основного тона.

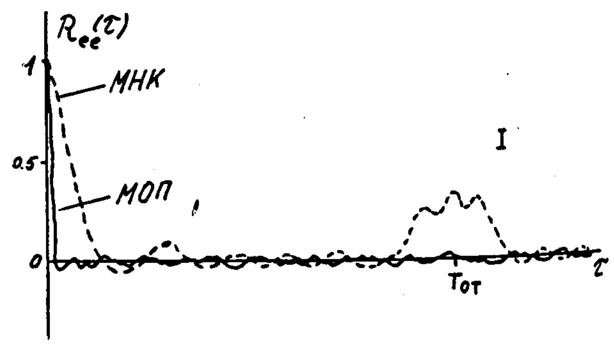


Рис.2 Корреляционные функции ошибки.

4. Разработка вокодера.

На основе алгоритма (8) на ЭВМ проведено моделирование цифрового вокодера с улучшенным качеством звучания восстановленного речевого сигнала в акустических шумах умеренной интенсивности SNR=(+5...+10дБ). Скорость передачи 2400 бит/с.

В качестве исходного материала был использован речевой сигнал с полосой частот до 4;7 кГц при частотах дискретизации 8;16 кГц соответственно и числе уровней квантования 2^{12}.

Блок-схема передающей части (анализатора) вокодера представлена на рис.3.

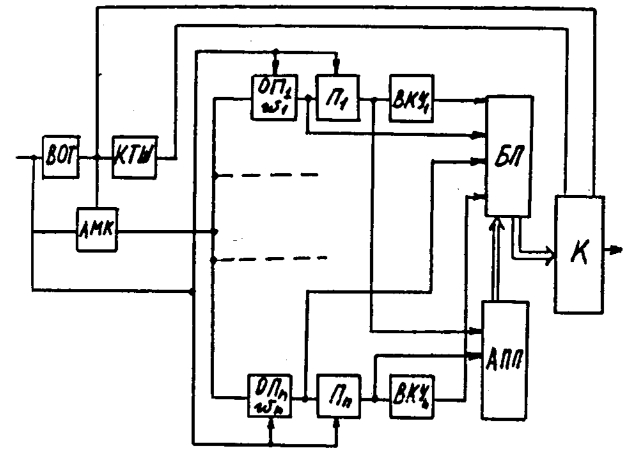


Рис.3 Блок-схема анализатора вокодера.

Передатчик состоит из:

блоков предварительной обработки (выделитель основного тона (ВОТ), классификатор тон-шум (КТШ), анализатор максимума корреляции (АМК), осуществляющих оценку периода основного тона, признака вокализованности, а также поиска точки взвешивания оценки корреляционной функции текущего сегмента сигнала;

пяти ветвей анализа (блок оценивания параметров ОП), блок предсказания (П), блок вычисления коэффициентов усиления (БКУ) с различными значениями функции веса $w(\tau)$;

блока памяти, в который записываются реализации погрешности предсказания и коэффициенты усиления;

блока анализа погрешности предсказания (АПП), который осуществляет выбор номера ветви по минимальному расстоянию между оценкой функции корреляции погрешности предсказания для данной ветви и функцией корреляции порождающего процесса;

блока квантования (К).

Экспериментальные исследования вокодера показали следующее:

I) При SNR = 5...10 дБ разборчивость слов равна 97%.

2) Улучшение качества синтезированного сигнала по сравнению с МНК достигается за счет уточнения спектров сигнала на переходных участках речи, которые плохо воспроизводятся в традиционных вокодерах. Близость спектров или корреляционных функций исходного и синтезированного сигналов в разработанном вокодере улучшается на 15...20%.

В таблице I приведены количественные соотношения для квадратичных отклонений функций исходного и синтезированного сигналов при различных SNR. Ошибка ε_R^2 равна нормированному квадрату нормы разности корреляционных функций исходного и синтезированного сигналов на периоде основного тона.

Таблица I.

SNR (дБ)	∞	I2	8
ε_R^2 (МНК)	0,4I±0,04	0,45±0,03	0,53±0,0I
ε_R^2 (МОП)	0,307±0,03	0,37±0,04	0,48±0,0I

3) Методом парных сравнений установлено, что число действительных суждений, высказанных аудиторами в пользу разработанного вокодера составило 80...85%.

Литература

I. Прохоров Ю.Н. Статистические модели и рекурентное предсказание речевых сигналов.-М.:Радио и связь,1984.-240с.

2. Лихачев С.Ф. Нелинейное предсказание речевых сигналов.-Материалы Всесоюзн. семинара АРСО-I2.-Киев,1982,с.II2-II4.

3. Назаров М.В.,Прохоров Ю.Н. Методы цифровой обработки и передачи речевых сигналов.-М.:Радио и связь,1985.-I76с.

4. Jain V.K. Speech signal analysis by error-weighted LPC-GIOBECOM'82: IEEE Global Telecommun. conf. Miami Beach, Fla,29 Nov-2 Dec, 1982, Conf.Rec.Vol 3,

New York, 1982, pp 1321-1324.
5. Un C.K., Lee J.R. On spectral flattening techniques in residual - exited linear precliction vocoding - ICASSP' 82, Proc. IEEE; Iht, Conf.Acoust. Speech and Signal Proc, Paris, May 3-5, 1982,NY, pp 216-219.

ON UNIVERSAL AND SPECIFIC FEATURES IN VOWEL PERCEPTION

K.S. OGORODNIKOVA

Dept.of Phonetics
Leningrad State Univ.
Leningrad, USSR, 199164

ABSTRACT

The results of a cross-language study of the perception of a set of synthetic steady-state vocalic stimuli using mimicking and identification methods are reported. The subjects were native speakers of Russian, French and Georgian. The results show the influence of the vocalic system of the mother tongue on vowel perception. A close correlate to the given stimulus occuring in the native vowel system induces significant changes in mimicking and identification responses. This influence may be manifest even in cases where this correlate is a context-bound allophone. A superficial aquaintance with the vocalic system of a second language changes the identification results, which has implications for the analysis of experimental perceptual data.

INTRODUCTION

The present paper attempts to establish, to what extent vowel perception of different language speakers is determined by the vowel system of their mother tongue, and to what extent - by the universal perceptual abilities of human listeners.

A number of researchers have maintained that speakers of different languages are able to identify more vowels than the number of vowel phonemes in the language they are speaking. However, neither a finite inventory of such perceptual vowel units, nor their relation to linguistic phonemes has as yet been established for any language.

Two possible solutions have been suggested for native speakers of Russian: I) this set of internal vowel representations might correspond to context-bound allophones in Russian vowels /3,2/; 2) it might conform to cardinal vowels /I/. But these solutions are not fully supported by the actual experimental data in different perceptual tests.

A combination of mimicking and identification was used. There is evidence to believe that the transformations of the initial signal in mimicking and identification coincide up to the phonetic feature level. In mimicking, transformation of the phonetic representation into motor commands then takes place. Identification requires the phonetic labelling step. Mimicking does not seem to imply a necessary phonemic classification, and when it is difficult, no decision in terms of phonological categories is made. The comparison of mimicking and identification results makes it possible to isolate motor and labelling factors.

It is important to realize that in analysing mimicking data purely in terms of FI and F2 values we lose a great deal of information about the phonetic quality of vowel responses.

PROCEDURE

Three groups of IO male adult subjects, native speakers of Russian, French and Georgian, took part in the experiments. A set of 8 synthetic steady-state vocalic stimuli with Fo increasing from IOO Hz to I25 Hz was used (phonetic symbols with a letter "s" are assigned to each stimulus).

Formant frequencies of synthetic vocalic stimuli

Stimuli	FI	F2	F3	F4
i_s	260	2760	2930	3500
y_s	240	I880	2660	3500
$ø_s$	350	I560	2200	3250
$æ_s$	840	I7I0	2200	3250
u_s	240	660	2420	3250
o_s*	290	600	2420	3250
$ɔ_s$	570	800	2420	3250
a_s	760	I060	3220	4000

The stimuli were recorded in random order at 5 ms interval, each stimulus was repeated 5 times.

There are eleven oral vowels in French: /i, e, ɛ, a, y, ø, œ, u, o, ɔ, ɑ/; six in Russian /i, e, ы, u, o, a/; five in Georgian: /i, e, a, u, o/. The Russian and the Georgian vowel systems are considerably poorer than the French one. On

*o_s-closer quality, $ɔ_s$-more open quality.

the other hand, large allophonic variations occur in Russian, unlike French and Georgian. Vowels differ in quality according to stress position and to the phonological palatalization of adjacent consonants (i-glides and an advanced vowel articulation).

The stimuli i_s, a_s, u_s have correlates in all three languages; y_s and $ø_s$ - front labial vowels - occur only in French. The stimuli $æ_s$, $ɔ_s$, $ρ_s$ have no close correlates in any of the three languages. However, $æ_s$ is phonetically nearer to the French /a/ and /ɛ/, $ɔ_s$ - to /ɔ/, $ρ_s$ - to /o/, than to Russian or Georgian vowels.

MIMICKING TEST

All the subjects were instructed to repeat as closely as possible the stimuli that they thought to be natural. Each subject went through the mimicking test twice and gave IO responses to each stimulus, which were recorded onto tape. Before mimicking, subjects pronounced vowels in their own language.

The FI and F2 values of the response vowels were measured from spectrograms and plotted as dots on the FI/F2 plane. The accumulations of such dots formed the response areas for each stimulus by each group (see Fig.I a,b,c).

All the vowel responses were classified using phonetic symbols and signs for finer phonetic details by a trained phonetician (see Table I for the results). The response areas to different vocalic stimuli partly overlap, less in the case of French speakers and most of all in the case of Georgian speakers.

All the subjects responded to i_s, a_s, u_s stimuli with their own corresponding vowels.

Only French subjects were successful in mimicking y_s and $ø_s$. Russian subjects showed much poorer results and those of Georgian subjects were on the whole inadequate.

French and Russian subjects gave similar responses to $æ_s$; the Georgians responded by an /a/, often pharyngalized.

Russian and Georgian subjects tended to substitute their own vowels for $ɔ_s$ and $ρ_s$ stimuli. French subjects' responses were sometimes phonetically rather close to $ɔ_s$ and $ρ_s$.

Thus, mimicking results were strongly determined by the linguistic experience of the subjects: mimicking is more accurate when the stimulus has a correlate in the vocalic system of the mother tongue. It was therefore to be expected that the mimicking of French subjects would be most accurate.

But a vowel without correlates in the subjects' native language can also be accurately responded to. The better mimicking results of Russian subjects in comparison with Georgian ones seem to be due to the advanced articulation of the

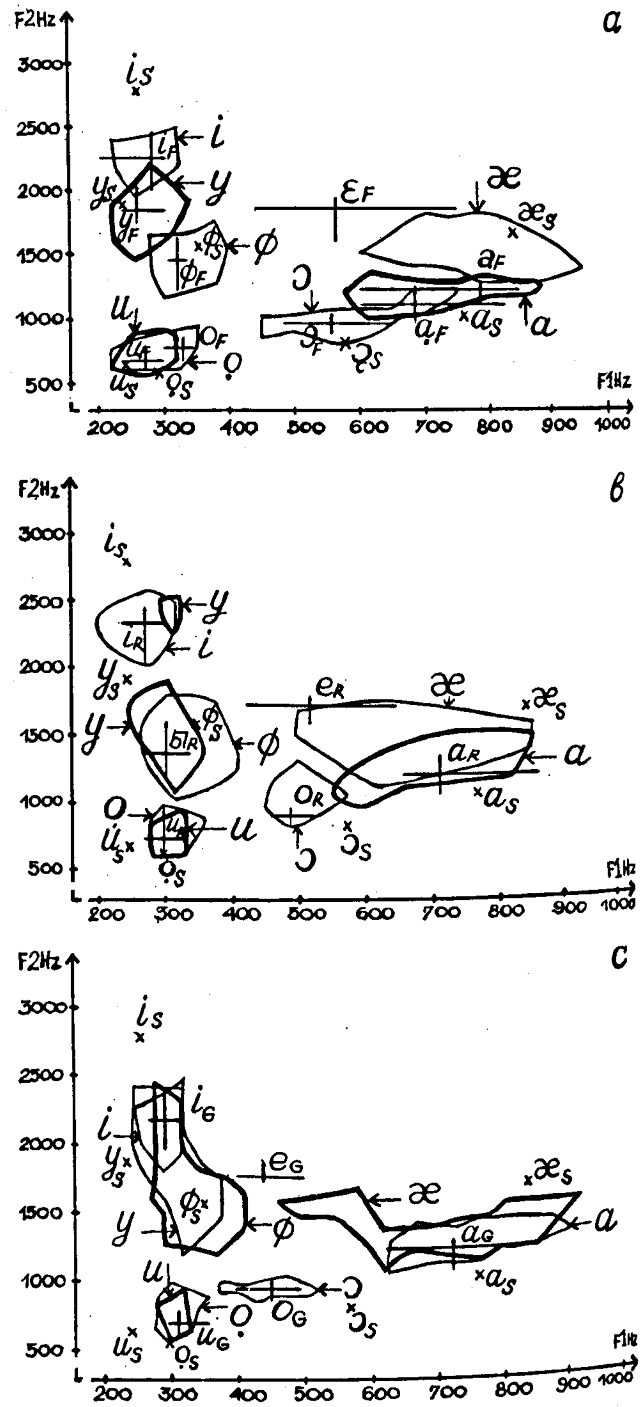


Fig.I a,b,c. Mimicking response areas in the FI/F2 plane of French (a), Russian (b) and Georgian (c) subjects. Dots with symbols $ə_F$, a_F etc. show the locations of French vowels, a_R, i_R etc.- Russian and a_G, i_G - Georgian vowels (the mean of 20 measurements, mini and maxi FI and F2 values are also shown).

Russian /u, o, a/ adjacent to the palatalized consonants. The perceptual independence of such allophones is reinforced by the existence of special letters for them in the Russian alphabet.

		i_s	y_s	$ø_s$	$æ_s$	a_s	$ɔ_s$	$ρ_s$	u_s
i	I	I00						I	2
	2	60							
	3	67	36	I5					
i/ɪ	I								
	2	34	9						
	3	29	3I	20					
ɪ	I		I						I
	2	5	I6	25					
	3	3	8	26					
ы	I								
	2		35	24					
	3		I2	I6					
y	I		97						
	2		28	5					I
	3	I	I3	I2					
ø	I		2	I00					
	2		7	7					I
	3			7					
œ	I								
	2	I	4	22			2		
	3						7		
ə	I								
	2			I7	I				I
	3			I	2			I	
ɛ	I				29				
	2				37				
	3			3	6	I			
æ	I				I9				
	2				I4				
	3								
ɑ	I				47	8			
	2				33	28			
	3				4I	I4			
ɑ/a	I					2I			
	2				I0	26	I		
	3				II	24			
a	I				5	7I	28		
	2				5	46	8		
	3				40	6I			
ɒ	I						49		
	2						I4		
	3								
ɔ	I						2I		
	2						36		
	3						I2		
o	I							8	
	2						34		
	3						88		
o/u	I							36	I9
	2		I					3I	22
	3							32	20
u	I							55	78
	2							69	75
	3							67	80

Table I. Mimicking responses to synthetic vowels by groups of French /I/, Russian /2/ and Georgian /3/ speakers.

Russian, and only rarely Georgian speakers responded to y_s and $ø_s$ with the unrounded vowels: a central /ы/ or a retracted /ɪ/, realizing the same low values of F2 owing to vowel retraction and not to vowel rounding.

IDENTIFICATION TEST

The same subjects after a delay of several days were instructed to provide a possibly exact graphical representation of the same set of stimuli as in the mimicking test. See the results of the classification of the diverse responses in Table 2.

Identification and mimicking responses to each stimulus by the three groups of subjects have much in common: the best results in the three groups were for i_s, a_s, u_s vowels; the most adequate responses were from French subjects; there were better responses from Russian than from Georgian subjects to y_s, $ø_s$, $æ_s$ stimuli etc. It should be specially noted, that a Russian subject identified $æ_s$ as /'a/ - after a palatalized consonant, while a Georgian one - as /ħ/, that is, he perceived pharyngalization and considered it to be the most prominent feature.

A supplementary test was conceived to verify our assumption that even a superficial aquaintance with the vocalic system of a second language may influence the perception of vowels that do not occur in the mother tongue as context-free allophones. The identification of the same set of vocalic stimuli was tested with a group of native speakers of Georgian, all - first year students in physics at Tbilisi University. Of the total of 38 subjects - I6 had studied English as a foreign language at school and 22 - French and German. It was found that those who had studied English did not respond to y_s with rounded vowels at all and gave almost no responses to $ø_s$ with a front rounded vowel. Those who had studied French and German identified y_s as a rounded vowel and $ø_s$ - as a front rounded vowel in I/3 of their responses.

Thus, the results of mimicking and of identification of vocalic stimuli proved to be similar, but mimicking was still more accurate: the subjects responded with similar vowel types in both tests. In general, this is also true for each individual subject. Sometimes, however, subjects answered with different vowels from test to test: for example, mimicking responses to $ɔ_s$ as /a/ and identification responses as /ɔ/ of a French subject.

If mimicking responses were influenced by individual articulation skill, the identification responses even in a free-choice experimental situation were to a great extent determined by the subjects' resourcefulness in choosing an appropriate symbol. (For example, a Georgian identified y_s as "fi" - the consonant

	is	ys	øs	æs	as	ɔs	os	us
i	I00 98 92	4 72	56					
i/ы		I0	I2 6					
ы	2	30 I0	28 I8					
y		I00 50 I0	2 I0					
ø			78 48					
œ			22					
ə	6	8	4	2				
e/ε			I0 4	30 I4 20				
a/ε				28 40 2				
h			2	I0				
'a				8				
a				36	60			
a/a				6 34 68	6 I00 I00			
a					34	I0		
ɔ/a						I8		
ɔ/ə				2		72 I6 I0		
o						84 90	8	
u/o							I8 8	2
u	2	6					74 92 I00	98 I00 I00

Table 2. Identification responses to synthetic vowels by groups of French (top figure for each classification unit), Russian (middle figure) and Georgian (bottom figure) speakers.

seemingly carried the feature of "lip articulation", i.e. "rounded").

CONCLUSION

The results reported above suggest the influence of the vocalic system of the mother tongue on vowel perception. The set of synthetic vowels was most compatible with the linguistic experience of the French subjects, and they had the best results in identification.
But this influence is more complex than the presence of a close corresponding vowel to the stimulus in the vocalic system. We may assume a certain role of acoustical properties of the native vowels involved as references in the perceptual process.
Furthermore, we may speculate that not only the phonetic properties of the context-free allophones, but also of the most perceptually distinct context-bound allophones of the native vowels exert a certain influence on vowel perception. The better results in mimicking and identification of ys, øs, æs achieved by native speakers of Russian than by the Georgians seem to be due to the actual advancement of the Russian /u, o, a/ allophones adjacent to palatalized consonants.
On the contrary, large allophonic variations do not occur in the Georgian language and Georgian subjects tend to give more "categorical" responses.
The obtained results cannot be explained only by the influence of the phonological system of the mother tongue, but also reflect the universal perceptual abilities of different language users.
And finally, it is suggested that even a superficial aquaintance with the vowel system of a second language has an effect on vowel perception which should be borne in mind when interpreting the results of perceptual experiments.

REFERENCES

/I/ Chernova E.I., Beliakova G.A., Malinnikova T.G. A Study in Perception of Cardinal Vowels by Native Speakers of Russian. - Z.Phon.Sprachwiss.Kommunik.forsch., 39(4), Berlin, I986, 472-483 /in Russian/.
/2/ Chistovich L., Fant G., Serpa-Leitão A., Tjernlund P. Mimicking of Synthetic Vowels. - Speech Transm. Lab.Quart.Progr. and Status Report, Stockholm, I966, N 2, I-I8.
/3/ Verbitskaya L.A. On Perceptual Vowel Units in Russian Speech. - Ucheniye Zapisky Leningradskogo Universiteta, 325(69), Leningrad, I964, 55-7I /in Russian/.

CATEGORICAL PITCH PERCEPTION

KLAUS J. KOHLER

Institut für Phonetik und
digitale Sprachverarbeitung
Universität Kiel
2300 Kiel, FRG

ABSTRACT

This paper shows that the paradigm of categorical perception also applies to pitch contours. In LPC-synthesized stimuli, an F0 peak is shifted through an utterance in 30-ms steps. The stimuli of this physical continuum are identified in a contextualization experiment. The response function shows an abrupt change when the F0 peak is moved into the vowel of the stressed syllable. When stimuli from the continuum are paired with 0, 1 or 2 steps between them, the differentiation functions show maxima at the category boundary established by the identification test. The ordering in each pair has an influence on the differentiation function.

INTRODUCTION

The paradigm of categorical perception is well-known in the area of sound segments /1/. It means that a physical continuum of a sound property is partitioned into sections inside which the same category is identified and between which category identification changes. The corollary of this is that differentiation along the physical continuum is sharpest across the category boundaries and weakest inside the categories. The evidence for this phenomenon in the perception of prosodic features, e.g. word tones in tone languages, is contradictory /1,2/, and it certainly has not been demonstrated for utterance pitch contours. To show its relevance in the field of intonation the following experiments were carried out.

PROCEDURE

In the German sentence "Sie hat ja gelogen." ("She's been lying."), with focus stress on the syllable "-lo-" /lo:/, the F0 peak can be on the syllable "ge-", preceding the stress, or at the centre of the stressed syllable, or at its end (cf. /3/). This shift in the F0 peak position is correlated with a change in meaning from 'established' to 'new' to 'emphatic'. A token of this sentence was pronounced by a male speaker, LPC-analyzed, and resynthesized with 11 F0 contours, in which the peak was shifted in 30-ms steps from "ge-" to "-en" (for further details cf. /3/).

Experiment 1.
The first 8 stimuli out of this series of 11 (counting from left to right) were each paired with the preceding context "Jetzt versteh ich das erst." ("Now I understand."; spoken by the same speaker, and LPC-resynthesized). This precursor sets a semantic frame of reference for something new to follow in the test utterance. Since the 8 test stimuli span the continuum of F0 peak positions from "ge-" to the centre of the stressed syllable "-lo-", they either contain the same semantic component as suggested by the context frame, i.e. 'new', or the different meaning feature 'established', which would be appropriate as a summing-up at the end of a chain of arguments, for instance after "Once a lyer, always a lyer; this also applies to Anne: ...". Thus the chosen context and each of the 8 test stimuli either form a semantic match or they do not. A test tape was prepared with a randomization of 80 pairings of context and test stimuli (8 stimuli x 10 repetions) and presented to 19 listeners who had to indicate on prepared answer sheets whether context and test sentence were semantically congruous.

Experiment 2.
Stimuli from the series of 11 were paired in such a way that they differed by 0, 1, or 2 steps of F0 peak position. All 1- and 2-step combinations were formed in both orders (2x10 and 2x9, respectively), and supplemented by identical stimulus pairings at the uneven rank numbers in the series (6). Two test tapes were prepared: (I) for the ascending rank order in stimulus pairs (i.e. left-to-right shift of the F0 peak), and (II) for the descending rank order (i.e. right-to-left shift). For each test tape, the 6 identical stimulus pairs were added; the resulting 25 pairs were then repeated once and randomized.

A group of 39 subjects listened to test tape (I), a different group of 34 subjects to test tape (II). Listeners had to indicate on prepared answer sheets whether they perceived a difference between the members of a pair.

RESULTS AND DISCUSSION

Figure 1 gives the identification function for Experiment 1: it shows an abrupt change from "matching" to "non-matching" judgments in spite of the gradual change along the physical continuum, and is thus clearly categorical. The answers "matching" or "non-matching", respectively, can be interpreted as the identification of two sentence meanings: 'summing-up conclusion' (A) versus 'new point of argumentation' (B). Stimuli 1-4 represent semantic category (A), stimuli 6-8 category (B); stimulus 5 is on the border between the two. The latter is characterized acoustically by being the first stimulus in the whole series (from left to right) that has the F0 peak in the stressed vowel /o:/: approximately 30 ms after vowel onset. In the stimuli 1-4, the F0 peak precedes the stressed vowel, and there is thus only an F0 fall in it; in the subsequent stimuli, the F0 fall in the stressed vowel is prefixed by a rise of increasing extent, which at a peak position of 60 ms into the vowel has become prominent enough to signal a different category in an identification task. We thus have a time span of about 60 ms into the vowel where the F0 peak is in a boundary area between two categories, and therefore has an equivocal meaning attached to it.

Figures 2a-c provide the discrimination functions for Experiment 2. The pairs of identical stimuli show a maximum of false alarms at the category boundary found in identification, i.e. for stimulus 5. This is what one would expect if the associated meaning is equivocal: listeners overdifferentiate at the perceptual level when the semantic attribution is unclear. In the pairs of different stimuli in the ascending order, the maximum of discrimination occurs at the category boundary of the identification function, as long as one member lies outside the transition span, i.e. for the pairings 4/5, 5/6; 3/5, 4/6, 5/7. This pattern changes in the pairs with descending order; the maximum is generally shifted to the next higher rank in the stimulus series: 6/5, 7/6; 6/4, 7/5, 8/6. This finding may be related to an upward shift of the transition span, the uncertain boundary area now being around stimulus 6. Such a boundary shift can be explained by perceptual hysteresis under the special conditions of the discrimination test paradigm.

If in a sequence of two segmentally identical utterances, i.e. a repetition of the same word string, two different F0 peak positions are selected from around the category transition as established by the identification test, the listener expects a descending F0 peak order, linked to a semantic shift from the category 'new' to the category 'established', as the unmarked case; a reversal of this order constitutes the marked case in this test frame because the repetition of the sentence suggests the progression from 'new' to 'established'. In this situation, perception becomes less acute to a decrease in the extent of a rising F0 (preceding the fall in the stressed vowel) than to an increase: the category boundary is raised in a right-to-left sequence of peaks, compared to its position determined by identification. Thus, stimulus 5, which lies between the two categories in the identification test, and which seems to stay there in left-to-right discrimination, is incorporated in the 'established' category in the reversed-order discrimination.

The maximum of differentiation between stimuli from an F0 peak position continuum is thus at the transition between identification categories. Therefore, the phenomenon of categorical perception also applies to the field of prosody, in particular to global utterance intonation. At the same time, however, a strong order effect which results from the perceptual testing procedures and which disturbs the differentiation functions has to be taken into account. It is found in segment perception, too, but has largely passed unnoticed because it has not been factored out.

REFERENCES

/1/ B.H. Repp, "Categorical perception: issues, methods, findings". In: N. J. Lass (ed.), Speech and Language: Advances in Basic Research and Practice. Vol. 10, pp. 244-335. Academic Press: Orlando, 1984.

/2/ A.S. Abramson, "The noncategorical perception of tone categories in Thai," In: B. Lindblom & S. Öhman (eds.), Frontiers of Speech Communication Research, pp. 127-134. Academic Press: New York, 1979.

/3/ K.J. Kohler, "Computer synthesis of intonation", Proc. 12th Intern. Congr. Acoustics, A6-6. Toronto, 1986.

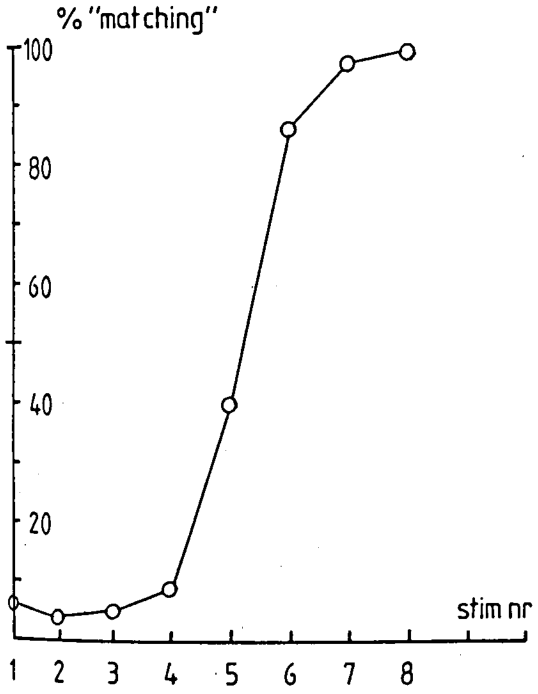
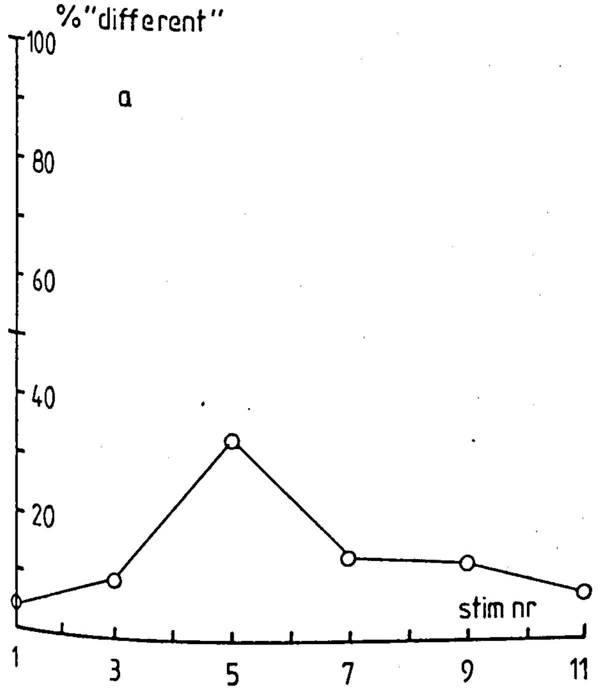
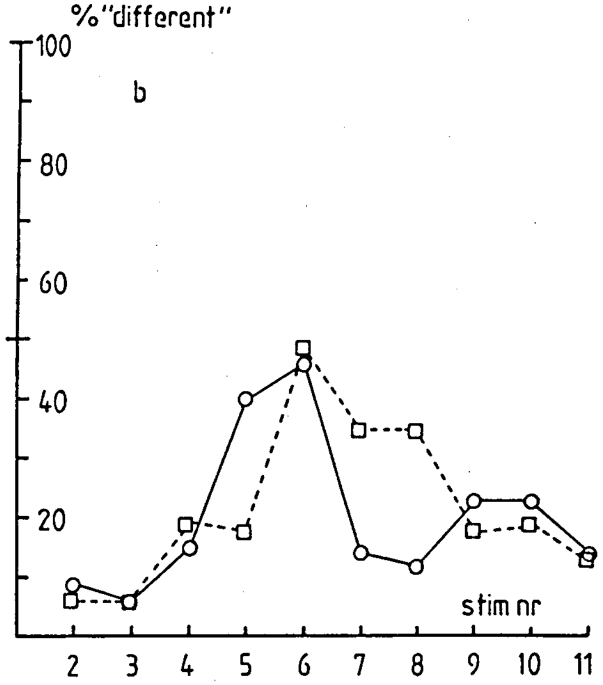


Fig. 1. Identification function in Experiment 1, showing percentage of "matching" judgements for 8 stimuli "Sie hat ja gelogen." with F0 peak shift from left to right in the context "Jetzt versteh ich das erst." 19 subjects; for each stimulus N=190.





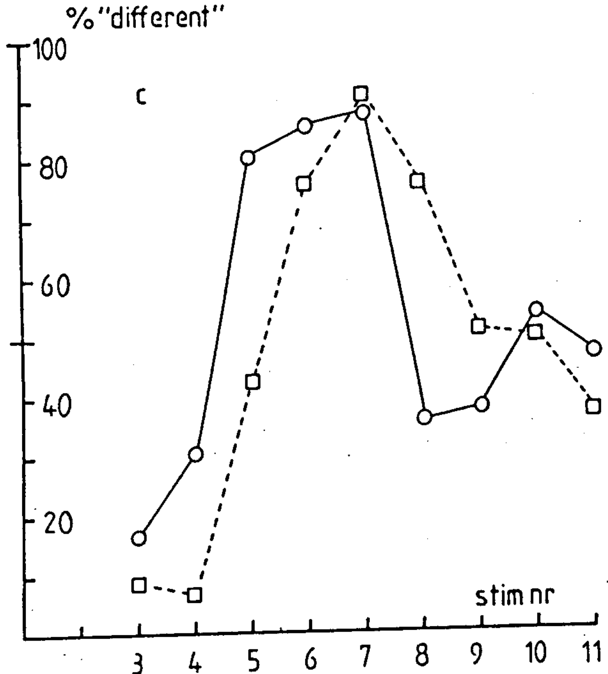


Fig. 2. Discrimination functions in Experiment 2, showing percentage of "different" judgements for utterance pairs of "Sie hat ja gelogen." with 0-(a), 1-(b), or 2-step (c) distances of F0 peak positions, in the ordering left-to-right (continuous line) or right-to-left (broken line). The stimulus numbers refer to the second stimulus in the ascending and to the first in the descending order. 73 sbs., N=146 at each data point (a); 39 sbs., N=78 in the left-to-right, 34 sbs., N=68 in the right-to-left ordering of (b) and (c).

A FUZZY LOGICAL MODEL OF SPEECH PERCEPTION

DOMINIC W. MASSARO

Program in Experimental Psycholgy, University of California,
Santa Cruz, California 95064 U.S.A.

ABSTRACT

Speech perception is viewed as having available multiple sources of information supporting the identification and interpretation of the language input. The results from a wide variety of experiments can be described within a framework of a fuzzy logical model of perception. The assumptions central to the model are 1) each source of information is evaluated to give the degree to which that source specifies various alternatives, 2) the sources of information are evaluated independently of one another, 3) the sources are integrated to provide an overall degree of support for each alternative, and 4) perceptual identification and interpretation follows the relative degree of support among the alternatives. A formalization of these assumptions is applied to results of an experiment manipulating audible and visible characteristics of the syllables /ba/ and /da/. In addition, the results are used to test an alternative categorical model of speech perception. The good description of the results by the fuzzy logical models indicate that the sources of support provide continuous rather than categorical information. The integration of the multiple sources results in the least ambiguous sources having the most impact on processing. These results provides major constraints to be met by theories of speech perception and language processing.

INTRODUCTION

Speech perception is a human skill that rivals our other impressive achievements. Even after decades of intense effort, speech recognition by machine remains far inferior to human performance. The central thesis of the present proposal is that there are multiple sources of information supporting speech perception, and the perceiver evaluates and integrates all of these sources to achieve perceptual recognition. Consider recognition of the word *performance* in the spoken sentence

The actress was praised for her outstanding performance.

Recognition of the critical word is achieved via a variety of bottom-up and top-down sources of information. Top-down sources include semantic, syntactic, and phonological constraints and bottom-up sources include audible and visible features of the spoken word.

A THEORETICAL FRAMEWORK FOR PATTERN RECOGNITION

According to the present framework, well-learned patterns are recognized in accordance with a general algorithm, regardless of the modality or particular nature of the patterns [1, 2, 3]. The model has received support in a wide variety of domains and consists of three operations in perceptual (primary) recognition: feature evaluation, feature integration, and pattern classification. Continuously-valued features are evaluated, integrated, and matched against prototype descriptions in memory, and an identification decision is made on the basis of the relative goodness of match of the stimulus information with the relevant prototype descriptions. The model is called a fuzzy logical model of perception (abbreviated FLMP).

Central to the FLMP are summary descriptions of the perceptual units of the language. These summary descriptions are called prototypes and they contain a conjunction of various properties called features. A prototype is a category and the features of the prototype correspond to the ideal values that an exemplar should have if it is a member of that category. The exact form of the representation of these properties is not known and may never be known. However, the memory representation must be compatible with the sensory representation resulting from the transduction of the audible and visible speech. Compatibility is necessary because the two representations must be related to one another. To recognize the syllable /ba/, the perceiver must be able to relate the information provided by the syllable itself to some memory of the category /ba/.

Prototypes are generated for the task at hand. In speech perception, for example, we might envision activation of all prototypes corresponding to the perceptual units of the language being spoken. For ease of exposition, consider a speech signal representing a single perceptual unit, such as the syllable /ba/. The sensory systems transduce the physical event and make available various sources of information called features. During the first operation in the model, the features are evaluated in terms of the prototypes in memory. For each feature and for each prototype, featural evaluation provides information about the degree to which the feature in the speech signal matches the featural value of the prototype.

Given the necessarily large variety of features, it is necessary to have a common metric representing the degree of match of each feature. The syllable /ba/, for example, might have visible featural information related to the closing of the lips and audible information corresponding to the second and third formant transitions. These two features must share a common metric if they eventually are going to be related to one another. To serve this purpose, fuzzy truth values [4] are used because they provide a natural representation of the degree of match. Fuzzy truth values lie between zero and one, corresponding to a proposition being completely false and completely true. The value .5 corresponds to a completely ambiguous situation whereas .7 would be more true than false and so on. Fuzzy truth values, therefore, not only can represent continuous rather than just categorical information, they also can represent different kinds of information. Another advantage of fuzzy truth values is that they couch information in mathematical terms (or at least in a quantitative form). This allows the natural development of a quantitative description of the

phenomenon of interest.

Feature evaluation provides the degree to which each feature in the syllable matches the corresponding feature in each prototype in memory. The goal, of course, is to determine the overall goodness of match of each prototype with the syllable. All of the features are capable of contributing to this process and the second operation of the model is called feature integration. That is, the features (actually the degrees of matches) corresponding to each prototype are combined (or conjoined in logical terms). The outcome of feature integration consists of the degree to which each prototype matches the syllable. In the model, all features contribute to the final value, but with the property that the least ambiguous features have the most impact on the outcome.

The third operation during recognition processing is pattern classification. During this stage, the merit of each relevant prototype is evaluated relative to the sum of the merits of the other relevant prototypes. This relative goodness of match gives the proportion of times the syllable is identified as an instance of the prototype. The relative goodness of match could also be determined from a rating judgment indicating the degree to which the syllable matches the category. The pattern classification operation is modeled after Luce's [5] choice rule. In pandemonium-like terms [6], we might say that it is not how loud some demon is shouting but rather the relative loudness of that demon in the crowd of relevant demons. An important prediction of the model is that one feature has its greatest effect when a second feature is at its most ambiguous level. Thus, the most informative feature has the greatest impact on the judgment.

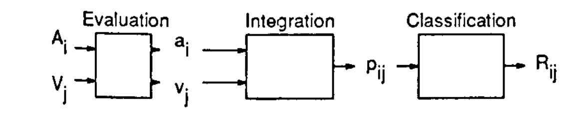


Figure 1. Schematic representation of the three operations involved in perceptual recognition.

Figure 1 illustrates the three stages involved in pattern recognition. Auditory and visual sources of information are represented by uppercase letters. The evaluation process transforms these into psychological values (indicated by lowercase letters) that are then integrated to give an overall value. The classification operation maps this value into some response, such as a discrete decision or a rating. The model confronts several important issues in describing speech perception. One issue has to do with whether multiple sources of information are evaluated in speech perception. Two other issues have to do with the evaluation of the sources in that we ask whether continuous information is available from each source and whether the output of evaluation of one source is contaminated by the other source. The issue of categorical versus continuous perception can also be asked with respect to the output of the integration process. Questions about integration assess whether the components passed on by evaluation are integrated into some higher-order representation and how the two sources of information are integrated.

The theoretical framework of the FLMP has proven to be a valuable framework for the study of speech perception. Experiments designed in this framework have provided important information concerning the sources of information in speech perception, and how these sources of information are processed to support speech perception. The experiments have studied a broad range of information sources, including bottom-up sources such as audible and visible characteristics of speech and top-down sources, including phonological, lexical, syntactic, and semantic constraints. As examples, experiments have assessed the contributions of formant structure and duration of vowels in vowel identification [7], the role of vowel duration and consonant duration in the identification of post-vocalic stop consonants [8, 9] and fricatives [10], the integration of voice onset time and formant structure of segment-initial stop consonants [11, 12] and fricatives [13]. These results are not limited to western languages; experiments have shown that both pitch height and pitch contour contribute to the perception of Mandarin Chinese lexical tone [14]. Experiments have also revealed the integration of nonauditory sources of information, such as pointing gestures, with auditory sources [15]. Several experiments have also addressed the relative contributions of acoustic information and higher-order constraints in the pattern. These experiments have included formant structure and phonological constraints in the identification of glides [16], the formant structure and lexical constraints in the identification of stop consonants [17], segmental information and syntactic constraints in the identification of words [18], semantic constraints in word identification [19], and word order, animacy, and noun-verb agreement in sentence interpretation [20].

EXPANDED FACTORIAL DESIGN

An expanded factorial design with open-ended response alternatives offers the potential of addressing important issues in speech perception. I will describe an experiment manipulating auditory and visual information in a speech perception task. The novel design illustrated in Figure 2, along with open-ended response alternatives, has not been used previously in speech perception research and it provides a unique method to address the issues of evaluation and integration of audible and visible information in speech perception.

Eight college students from the University of California, Santa Cruz, participated for one hour in the experiment. All test stimuli were recorded on videotape. On each trial the speaker said either /ba/ or /da/ or nothing, as cued by a video terminal under computer control. When the speaker was cued to say nothing, a

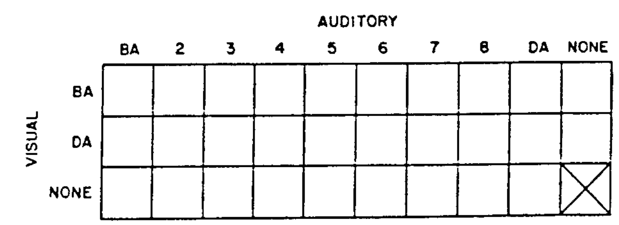


Figure 2.
Expansion of a typical factorial design to include auditory and visual conditions presented alone. The nine levels along the auditory continuum represent speech sounds varying in equal steps between /ba/ and /da/.

computer-controlled tone was recorded on the audio channel of the videotape 400 msec after the onset of the neutral cue. The original audio track of the videotape was replaced with synthetic speech. A nine-step /ba/ to /da/ auditory continuum was used to replace the original audio. By altering the parametric information specifying the first 80 msec of the consonant-vowel syllable, a set of nine 400 msec syllables covering the range from /ba/ to /da/ was created. The experimental videotapes were made by copying the original tape and replacing the original sound track with the synthetic speech. The presentation of the synthetic speech was synchronized with the original audio track on the videotape.

The 29 speech events illustrated in Figure 2 were presented to each subject in a randomized order. Each subject made about 600 identifications, which were converted to probabilities of responding with each of the eight alternatives. Figure 3 presents the observed probability of each of the eight responses for the 29 unique speech events.

FUZZY LOGICAL MODEL OF PERCEPTION (FLMP)

Applying the model to the present task using auditory and visual speech, both sources are assumed to provide continuous and independent evidence for the alternatives /ba/ and /da/. Defining the onsets of the second (F2) and third (F3) formants as the important auditory feature and the degree of initial opening of the lips as the important visual feature, the prototype for /da/ would be:

/da/ : Slightly falling F2-F3 & Open lips.

The prototype for /ba/ would be defined in an analogous fashion,

/ba/ : Rising F2-F3 & Closed lips,

and so on for the other response alternatives. Given a prototype's independent specifications for the auditory and visual sources, the value of one source cannot change the value of the other source at the prototype matching stage. The integration of the features defining each prototype is evaluated according to the product of the feature values. If aD_i represents the degree to which the auditory stimulus A_i supports the alternative /da/, that is, has Slightly falling F2-F3; and vD_j represents the degree to which the visual stimulus V_j supports the alternative /ba/, that is, has Open lips, then the outcome of prototype matching for /da/ would be:

/da/ : $aD_i\ vD_j$

where the subscripts i and j index the levels of the auditory and visual modalities, respectively. Analogously, if aB_i represents the degree to which the auditory stimulus A_i has Rising F2-F3 and vB_j represents the degree to which the visual stimulus V_j has Closed lips, the outcome of prototype matching for /ba/ would be:

/ba/ : $aB_i\ vBj$

and so on for the other alternatives.

The pattern classification operation would determine their relative merit leading to the prediction that

$$P(/da/\,|\,A_iV_j) = \frac{aD_i\ vD_j}{\sum} \qquad (1)$$

where $\sum$ is equal to the sum of the merit of all eight alternatives, derived in the manner illustrated for /da/ and /ba/.

The important assumption of the FLMP is that the auditory source supports each alternative to some degree and analogously for the visual source. Each alternative is defined by ideal values of the auditory and visual information. Each level of a source supports each alternative to differing degrees represented by feature values. The feature values representing the degree of support from the auditory and visual information for a given alternative are integrated following the multiplicative rule given by the FLMP. The model requires 2 parameters for the visual feature values and 9 parameters for the auditory feature values, for each of the 8 response alternatives, for a total of 88 parameters.

CATEGORICAL MODEL OF PERCEPTION (CMP)

It is essential to contrast one model with other models that make alternative assumptions. One alternative is a categorical model of perception (CMP). It assumes that only categorical

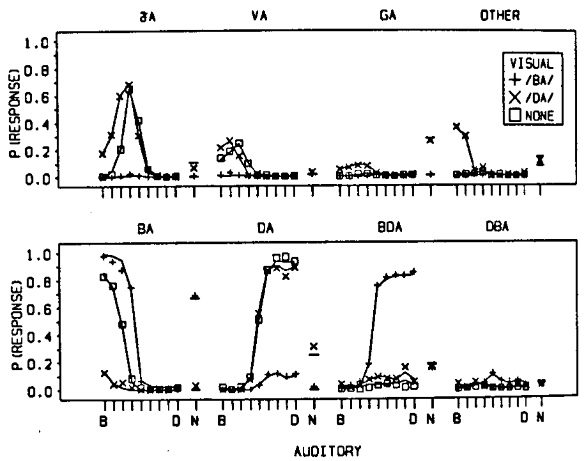


Figure 3. Probability of responding with each of the eight alternatives as a function of the auditory and visual sources under the bimodal and unimodal conditions. The nine levels between B and D along the auditory continuum represent speech sounds varying in equal steps between /ba/ and /da/. The level N refers to no auditory information. The curve parameter corresponds to a visual /ba/, a visual /da/, and no visual information. The lines give the predictions for the FLMP.

information is available from the auditory and visual sources and that the identification judgment is based on separate decisions to the auditory and visual sources. Given eight response alternatives, there are eight possible outcomes for a particular combination of auditory and visual information. Considering the /ba/ identification, the visual and auditory decisions could be /ba/-/ba/, /ba/-not /ba/, not /ba/-/ba/, or not /ba/-not /ba/. If the two decisions to a given speech event agree, the identification response can follow either source. When the two decisions disagree, it is assumed that the subject will respond with the decision to the auditory source on some proportion p of the trials, and with the decision to the visual source on the remainder $(1-p)$ of the trials. The weight p reflects the relative dominance of the auditory source.

The probability of a /ba/ identification response, P(/ba/), given a particular auditory/visual speech event, A_iV_j, would be:

$$P(/ba/\,|\,A_iV_j) = (1)\,aB_i\,vB_j + (p)\,aB_i\,(1-vB_j) + (1-p)(1-aB_i)vB_j + (0)(1-aB_i)(1-vB_j) \qquad (2)$$

where i and j index the levels of the auditory and visual modalities, respectively. The aB_i value represents the probability of a /ba/ decision given the auditory level i, and vB_j is the probability of a /ba/ decision given the visual level j. The value p reflects the bias to follow the auditory source. Each of the four terms in the equation represents the likelihood of one of the four possible outcomes multiplied by the probability of a /ba/ identification response given that outcome. To fit this model to the results, each unique level of the auditory stimulus requires a unique parameter aB_i, and analogously for vB_j. The modeling of /ba/ responses thus requires 9 auditory parameters plus 2 visual parameters. Each of the other seven response alternatives needs an analogous equation to Equation 2 and an additional 11 parameters, thus requiring a total of 88 visual and auditory parameters. For any particular auditory-visual combination, the sum of the eight decision probabilities to a given source also has to be constrained to be less than or equal to one; the assumption is that a given

source is categorized as only a single category on any given presentation. An additional p value would be fixed across all conditions for a total of 89 parameters. Thus, we have a fair comparison to the FLMP which requires 88 parameters.

MODEL TESTS

Figures 3 and 4 give the average observed results and the average predicted results of the FLMP and CMP. As can be seen in the Figure 4, the CMP gave a poor description of the observed results. The predictions of the FLMP shown in Figure 3, on the other hand, provide a very good description. The FLMP gave a mean root mean square deviation (RMSD) of .030 averaged across the individual subject fits of the 8 subjects compared to an average RMSD of .148 for the CMP.

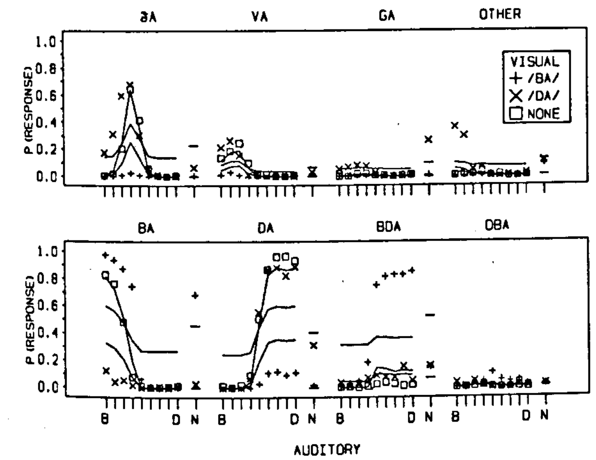


Figure 4. Probability of responding with each of the eight alternatives as a function of the auditory and visual sources under the bimodal and unimodal conditions. The nine levels between B and D along the auditory continuum represent speech sounds varying in equal steps between /ba/ and /da/. The level N refers to no auditory information. The curve parameter corresponds to visual /ba/, a visual /da/, and no visual information. The lines give the predictions for the CMP.

CONCLUSION

The present framework provides a valuable approach to the study of speech perception. We have learned about some of the fundamental stages of processing involved in speech perception by ear and eye, and how multiple sources of information are used in speech perception. Given the potential for evaluating and integrating multiple sources of information in speech perception and understanding, no single source should be considered necessary. There is now good evidence that perceivers have continuous information about the various sources of information, each source is evaluated, and all sources are integrated in speech perception. Future work should address the nature of the variety of sources of information, and how they function in recovering the speaker's message.

REFERENCES

[1] Massaro, D. W. (1979). Reading and listening (Tutorial paper). In P. A. Kolers, M. Wrolstad, & H. Bouma (Eds.), *Processing of visible language: Vol. 1* (pp. 331-354). New York: Plenum.

[2] Massaro, D. W., & Oden, G. C. (1980b). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice: Vol. 3* (pp. 129-165). New York: Academic Press.

[3] Massaro, D. W. (in press). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Lawrence Erlbaum Associates.

[4] Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8,* 338-353.

[5] Luce, R. D. (1959). *Individual choice behavior.* New York: Wiley.

[6] Selfridge, O. G. (1959). Pandemonium: A paradigm for learning. In *Mechanization of thought processes* (pp. 511-526). London: Her Majesty's Stationery Office.

[7] Massaro, D. W. (1984). Time's role for information, processing, and normalization. *Annals of the New York Academy of Sciences, Timing and Time Perception, 423,* 372-384.

[8] Denes, P. (1955). Effects of duration on the perception of voicing. *Journal of the Acoustical Society of America, 27,* 761-764.

[9] Massaro, D. W., & Cohen, M. M. (1983). Consonant/vowel ratio: An improbable cue in speech. *Perception & Psychophysics, 33,* 502-505.

[10] Massaro, D. W., & Cohen, M. M. (1977). The contribution of voice-onset time and fundamental frequency as cues to the /zi/-/si/ distinction. *Perception & Psychophysics, 22,* 373-382.

[11] Massaro, D. W., & Oden, G. C. (1980). Evaluation and integration of acoustic features in speech perception. *Journal of the Acoustical Society of America, 67,* 996-1013.

[12] Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85,* 172-191.

[13] Massaro, D. W., & Cohen, M. M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction, *Journal of the Acoustical Society of America, 60,* 704-717.

[14] Massaro, D. W., Cohen, M. M., & Tseng, C-Y. (1985). The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics, 13,* 267-289.

[15] Thompson, L. A., & Massaro, D. W. (1986). Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology, 42,* 144-168.

[16] Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception & Psychophysics, 34,* 338-348.

[17] Ganong, W. F. III. (1980) Phonetic categorization in auditory word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 110-125.

[18] Tyler, L. K., & Wessels, J. (1983). Quantifying contextual contributions to word-recognition processes. *Perception & Psychophysics, 34,* 409-420.

[19] Oden, G. C. (1978). Semantic constraints and judged preference for interpretations of ambiguous sentences. *Memory & Cognition, 6,* 26-37.

НЕЛИНЕЙНОСТЬ ПЕРИФЕРИЧЕСКОГО ОТДЕЛА ОРГАНА СЛУХА И ЯВЛЕНИЕ МАСКИРОВКИ

В.С. ШУПЛЯКОВ

Институт физиологии им. И.П.Павлова АН СССР
Ленинград

РЕЗЮМЕ

Работа посвящена выяснению вопроса о природе нелинейности эффекта маскировки. Показано, что с помощью экспериментов с маскировкой может быть выявлена только частотно-зависимая составляющая амплитудной характеристики слуховой системы. Предлагается гипотеза, что нелинейность маскировки обусловлена нелинейностью колебаний базилярной мембраны улитки внутреннего уха. Получены зависимости, позволяющие перейти от кривых маскировки к кривым возбуждения.

При исследовании механизмов восприятия сложных акустических сигналов, в том числе и речевых, значительное место занимают психоакустические эксперименты с маскировкой. Данные, полученные в этих экспериментах используются при оценке разрешающей способности органа слуха по частоте [I], при определении помехоустойчивости и разборчивости речи, в измерениях громкости [6]. Достоверность предлагаемых моделей обработки сигналов слуховой системой при восприятии в большой степени зависит от интерпретации тех кривых, которые измеряются в экспериментах с маскировкой и адекватности наших представлений относительно механизмов этого явления.

Одна из важнейших проблем, связанных с явлением маскировки заключается в том, что линейная зависимость между уровнем маскера и уровнем маскируемого тона сохраняется лишь при условии совпадения частот маскера и тест-тона; на всех других частотах тест-тона зависимость нелинейная рис. I.

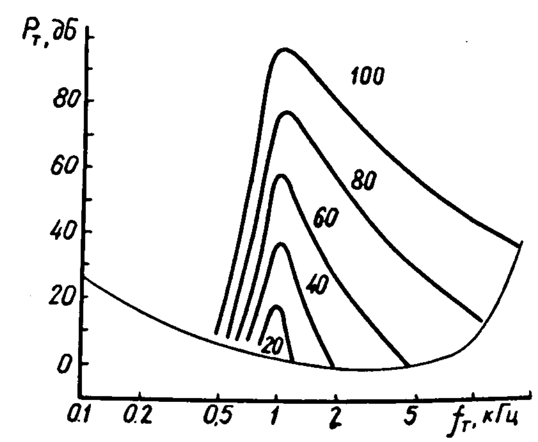


Рис. I. Кривые маскировки тона узкополосным шумом со средней частотой I кГц.

По оси абсцисс – частота тест-тона в кГц; по оси ординат – уровень звукового давления тест-тона в дБ.

Цифры у кривых – уровень маскера в дБ, штриховая линия – абсолютный порог слуха.

Начиная с классических опытов Вегеля и Лейна [5] возникновение нелинейных эффектов на высоких частотах объяснялось влиянием гармоник маскера, предположительно появляющихся вследствие нелинейности отделов слуха, предшествующих спектральному анализатору. В дальнейшем оказалось, что система наружного и среднего уха линейна в очень большом диапазоне уровней звуковых давлений, и таким образом, причины нелинейности маскировки остаются не выясненными до настоящего времени. Нелинейность кривых маскировки на частотах выше частоты маскера наблюдается и в экспериментах с задержанной маскировкой, а также при использовании в качестве маскера не только тональных, но и шумовых, речевых и других сигналов.

Для объяснения вышеупомянутого явления автор настоящего сообщения предлагает гипотезу, согласно которой нелинейность эффекта маскировки является отображением нелинейных свойств гидродинамической части улитки внутреннего уха и пытается определить вид этой нелинейности из экспериментов с маскировкой. Необходимо отметить, что представление о том, что кривые маскировки, в основном, описывают процессы, происходящие в периферическом отделе органа слуха, существует, начиная с самых первых экспериментов с маскировкой и широко принято в настоящее время. Считается, что кривые маскировки отображают характер распределения возбуждения, возникающего вдоль базилярной мембраны улитки внутреннего уха под воздействием входного сигнала. Такое представление лежит в основе современных расчетов громкости сигналов [6].

В данной работе будем рассматривать случай остаточной маскировки, чтобы упростить ситуацию и избежать необходимости учета не выясненного до настоящего времени механизма двухтонового подавления. При этом будем исходить из того, что картина распределения возбуждения, возникшая вдоль координаты X системы под действием маскера, сохраняется в течение некоторого времени после выключения маскера. Естественно предположить, что для восприятия в этих условиях тест-тона необходимо, чтобы амплитуда возбуждения, вызванного им на некоторой координате X_T была равна амплитуде возбуждения, вызванного маскером на этой координате, или отличалась от нее на некоторую небольшую величину ξ, представляющую собой, по существу, дифференциальный порог по интенсивности, т.е.

$$W_M(X_1 P_M) = W_T(X_1 P_T) + \xi(X_1 P_M) \quad (1)$$

Если величиной ξ пренебречь, поскольку она по крайней мере на порядок меньше точности самого метода маскировки, то получается, что какую бы форму амплитудной характеристики $W = W(P)$ не имела наша система, выявить её в экспериментах с маскировкой мы не сможем, поскольку и тест-тон и маскер подвергаются одному и тому же, в общем случае, нелинейному преобразованию. В результате экспериментов с маскировкой мы всегда будем получать линейную зависимость между уровнями маскера и тест-тона на всех частотах.

Поскольку, как уже было сказано выше, это противоречит экспериментальным данным, остается предположить, что амплитудные характеристики преобразования уровней звукового давления маскера и тест-тона в уровень возбуждения различны для маскера и тест-тона. Учитывая, что тест-тон отличается от маскера только частотой, остается принять, что амплитудная характеристика преобразования уровня звукового давления в уровень возбуждения частотно-зависима, т.е. каждая "координата" описывается не одной, а семейством амплитудных характеристик, зависящих от воздействующей частоты и специфичных для данной координаты.

Попытаемся теперь установить связь между кривыми, представленными на рис. I и амплитудными характеристиками системы. Пусть на рис. 2 схематически изображены две амплитудные характеристики $W_M(P)$ и $W_T(P)$ некоторой координаты X_I системы, соответствующие частотам маскера и тест-тона. Приращение уровня маскера на величину ΔP_M вызовет увеличение уровня возбуждения системы на некоторую величину λ. Для того, чтобы тест-тон был воспринят, необходимо увеличить его уровень на такую величину ΔP_T, которая приведет к увеличению уровня возбуждения на ту же величину.

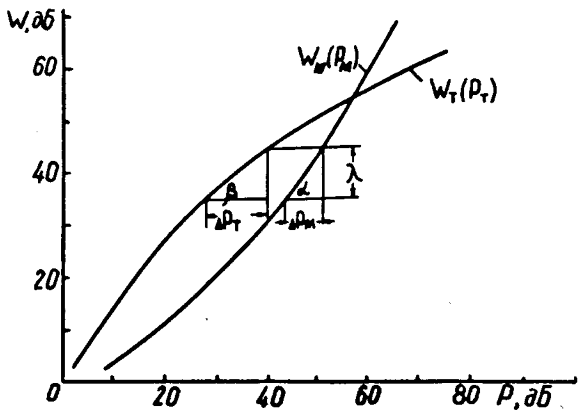


Рис. 2. Схематическое изображение двух амплитудных характеристик (для тест-тона и маскера) одной и той же координаты X_T базилярной мембраны.

По оси абсцисс – уровень звукового давления в дБ; по оси ординат – уровень возбуждения в дБ. Остальные пояснения в тексте.

Другими словами, можно записать:

$$\lambda = \Delta P_M \frac{\partial W_M(P_M)}{\partial P_M} = \Delta P_T \frac{\partial W_T(P_T)}{\partial P_T}$$

и отсюда

$$tg\gamma = \frac{tg\alpha}{tg\beta}$$

где γ – угол наклона характеристики $P_T(P_M)$ в точке X_I, а α и β – углы наклонов характеристик $W_M(P_M)$ и $W_T(P_T)$. Отсюда видно, что если, например, характеристики $W_M(P_M)$ и $W_T(P_T)$ совпадают, или просто одвинуты по горизонтали, $tg\gamma$ всегда будет равен I, т.е. мы получим прямые под углом 45°, что мы и наблюдаем в случае, если частоты маскера и тест-тона совпадают ОРис. 3).

Из выражения (3) также следует, что на основании экспериментов с маскировкой определить вид амплитудной характеристики системы на частоте тест-тона можно лишь при условии, если известна амплитудная характеристика системы на частоте маскера и наоборот.

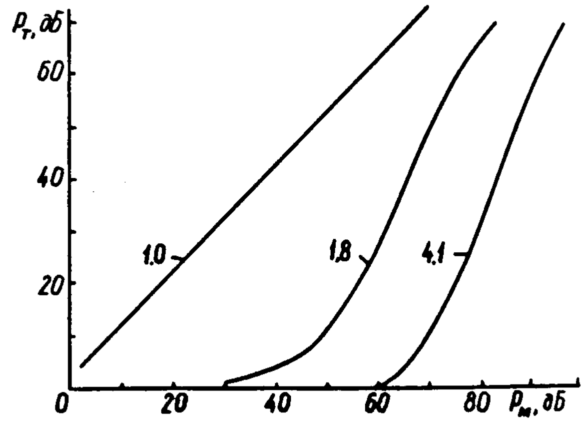


Рис. 3. Зависимость уровня звукового давления тест-тона от уровня звукового давления маскера, представляющего узкополосный шум со средней частотой I кГц.

По оси абсцисс – уровень звукового давления маскера в дБ; по оси ординат – уровень звукового давления тест-тона в дБ.

Цифры у кривых – частота тест-тона в кГц.

В настоящее время имеется достаточное число фактов, говорящих о том, что слуховая система уже на уровне механических колебаний базилярной мембраны улитки внутреннего уха колеблется нелинейно, причем эта нелинейность частотно-зависима [4] , [2] . Нелинейность этой амплитудной характеристики наиболее выражена на координатах мембраны, колеблющихся с резонансной частотой; на удаленных от резонанса частотах система линейна.

Естественно использовать этот факт для анализа амплитудных характеристик эффекта маскировки. В приложении к нашему случаю это означает, что если частота тест-тона достаточно удалена от частоты маскера, то амплитудная характеристика $W_M(P_M)$ на этой частоте линейна, т.е. $tg\alpha$ = I во всем диапазоне уровней маскера, а нелинейность кривых маскировки объясняется нелинейностью амплитудной характеристики на частоте тест-тона, т.е. $W_T(P_T)$. Тогда можно записать:

$$tg\gamma = \frac{1}{tg\beta}$$

Зависимость $P_T(P_M)$ может быть легко измерена (Рис. 3); зная эту характеристику, на основании (4) можно получить кривую зависимости $W_T(P_T)$ (Рис. 4).

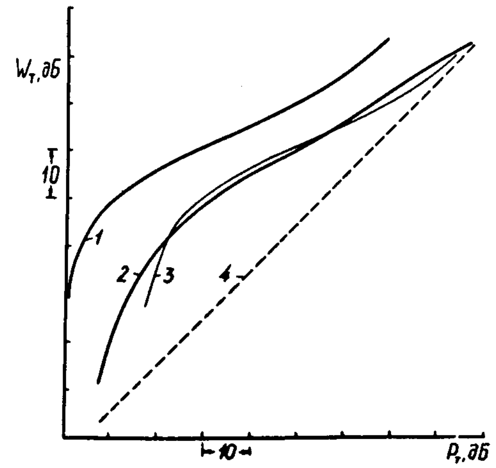


Рис. 4. Зависимость уровня возбуждения (кривые I и 3) и субъективной громкости тона I кГц (кривая 2) от уровня звукового давления тест-тона.

По оси абсцисс – уровень звукового давления тест-тона; по оси ординат – уровень возбуждения и субъективная громкость (по: I) в логарифмическом масштабе.

Кривая 3 получена путем сдвига кривой I; кривая 4 – прямая под углом 45°. (Характеристика $W_T(P_T)$ получается путем поворота на 90° зеркального изображения характеристики $P_T(P_M)$).

Как видно, эта кривая имеет три участка с характерными особенностями: при малых уровнях с ростом входного сигнала наблюдатеся непропорционально быстрое возрастание, "усиление" уровня возбуждения, что хорошо согласуется с современными представлениями об активном механизме формирования частотной избирательности периферического отдела слуха [3] . В области средних уровней наблюдается "насыщение" характеристики, это происходит за счет уменьшения избирательности (точнее, коэффициента неравномерности) частотно-избирательных кривых. При этом интересно отметить большое сходство этой кривой с кривой субъективной громкости (кривая 2 на рис. 4). Наконец, в области высоких уровней система становится все более линейной, что также хорошо согласуется с известными физиологическими экспериментами.

С учетом полученных результатов, характеристики распределения возбуждения вдоль базилярной мембраны приобретают совсем другой вид, чем это получается непосредственно из экспериментов по маскировке.

ЛИТЕРАТУРА

I. Цвикер Э., Р.Фельдкеллер. Ухо как приемник информации.Изд."Связь",Л.,I97I.
2. Шупляков В.С. Колебательные свойства структур улитки внутреннего уха. В сб. "Анализ сигналов на периферии слуховой системы", I98I.Изд."Наука", Л.,с.5-35.
3. Шупляков В.С. Математические модели гидродинамики улитки внутреннего уха. В сб. "Сенсорные системы.Слух", I982. Изд. "Наука", Л., с. 3-I7.
4. Rhode W.S., Robles L. Evidence from Mössbauer experiments for nonlinear vibration in the cochlea. – J.Acoust. Soc.Amer., 1974, v. 55, p. 588-596.
5. Wegel R.L., Lane C.E. The Auditory Masking of One Pure Tone by Another and its Probable Relation to the Dynamics of the Inner Ear. – Phys.Rev., 1924, v. 23, p. 266.
6. Zwicker E., Scharf B. A model of Loudness Summation. – Psychological Review, 1965, v. 72, N° 1, p. 3-26.

PERCEPTUAL AND ACOUSTICAL ANALYSES OF VELAR STOP CONSONANTS

SARAH HAWKINS

Department of Linguistics
University of Cambridge
Cambridge CB3 9DA UK

KENNETH N. STEVENS

Research Laboratory of Electronics and
Dept. of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge MA 02139 USA

Abstract

An acoustic property that distinguishes velar consonants from labial, alveolar, and dental consonants is a prominent midfrequency "compact" spectral peak, usually in the frequency range 800-4000 Hz. In a series of perceptual experiments, synthetic syllables with initial voiced and voiceless stop consonants were generated, and the spectral characteristics of the consonant burst were systematically manipulated to yield various degrees of prominence of a midfrequency spectral peak. From listener responses to these stimulii, we have determined that the property of compactness depends in part on the amplitude of the prominent spectral peak in relation to a peak at about the same frequency in the following vowel. Spectral analyses of a number of naturally spoken stop consonants in English have shown that the amplitude characteristics of the midfrequency spectral prominence of the burst are consistent with the perceptual data. However, the degree of prominence often shows fluctuations throughout the region encompassed by the burst and voicing onset in the following vowel.

1. Introduction

The most distinctive acoustic characteristic of velar stops is usually said to be a compact spectral prominence, in the midfrequency range of 800-4000 Hz. In Fig. 1 we see smoothed spectra of the burst and vowel onset of a naturally-spoken /ga/, together with the waveform. The burst spectrum has the classical compact midfrequency prominence. Another attribute of the pattern in Fig. 1 is that the amplitude of the spectral peak in the burst is comparable (within about 5 dB) to the amplitude of the corresponding spectral peak in the vowel. This characteristic of the burst in relation to the vowel is consistent with data reported by several investigators [1, 2]. Velar stops also have a number of secondary characteristics, such as bursts that are longer, and first-formant transitions that tend to be slower than those for bilabials and alveolars. Nevertheless, spectra of velar stops vary a great deal, and the concept of "compactness" is poorly understood.

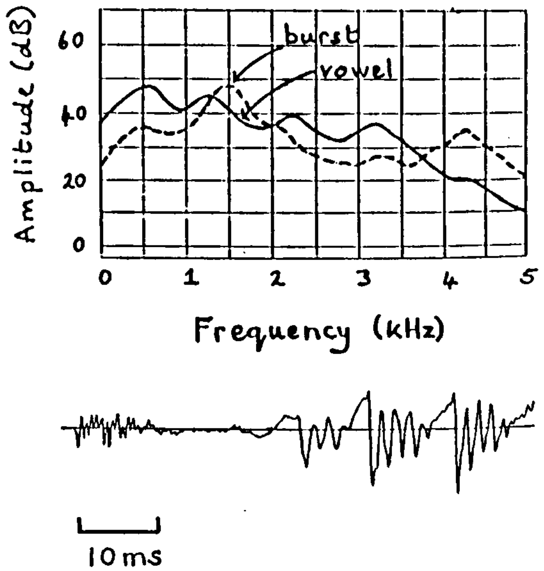


Fig. 1 Waveform (bottom) and spectra (top) sampled near the release of the syllable /ga/. Spectra are smoothed Fourier transforms sampled in the burst (dashed line) and at the onset of the vowel (solid line).

This paper describes some preliminary work in a planned series of studies of the acoustic characteristics of velar stops. We asked two questions, both of which focus on *spectral* rather than on temporal properties. First, can we synthesize an acceptable velar stop simply by manipulating the spectrum of the burst alone, and if so, what are the critical acoustic characteristics of such bursts? And second, to what extent are compact characteristics observable in *naturally*-spoken syllables? The focus of interest is the release burst and the first few milliseconds of the following vowel in syllable-initial stops.

2. Perceptual experiment

The stimuli for a perceptual experiment consisted of a series of synthetic consonant-vowel syllables. We constructed acoustic continua of bursts such that, when these bursts are followed by minimal vowels, we hear velar stops at one end of the continua, and either alveolars or bilabials at the other end, depending on the continuum. The various

bursts were synthesized manipulating the amplitudes of noise-excited formants in parallel synthesis.

Figure 2 shows short-time smoothed spectra of bursts at the extremes of the two continua—the velar-bilabial in the upper panel, and the velar-alveolar in the lower panel. In the velar-alveolar set of bursts, the classically compact shape of the spectrum labelled /g/ contrasts with the diffuse, rising spectrum, /d/, that is typical of alveolars. The variations in spectrum shape were achieved by changing the amplitude of excitation of $F2$. For the velar-bilabial set, formant amplitudes were altered so that the compact /g/ spectrum was made flatter and slightly falling, as for a bilabial.

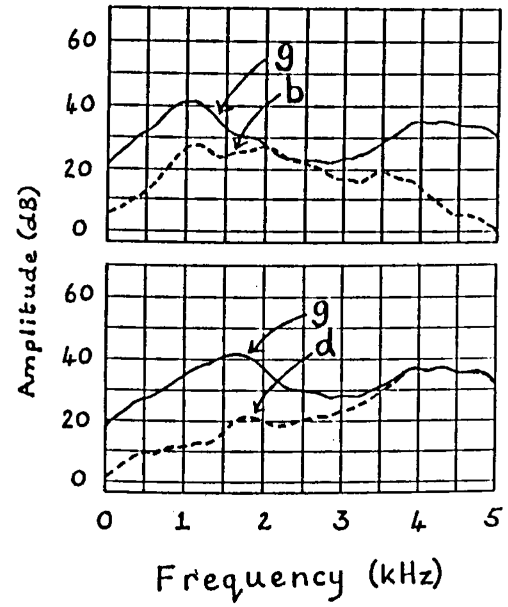


Fig. 2 Spectra of the bursts in the synthetic velar-bilabial continuum (top) and the velar-alveolar continuum (bottom). The spectrum at the velar end of each continuum is shown as the solid line and the spectrum at the bilabial or alveolar end is the dashed line.

Stimuli intermediate between these extreme pairs were made by changing the particular formant amplitudes in equal dB steps, resulting in two continua of 10 stimuli each. Listeners heard the bursts either in isolation, or with short, transitionless vowels following them, with or without aspiration. Since there were no transitions, the formant frequencies in the vowels were the same as those in the bursts (except for $F1$, for which there was a 20-ms rise at the onset from 250 to 500 Hz). $F2$ and $F3$ were lower in the velar-bilabial than in the velar-alveolar stimuli. For the velar-alveolar continuum, the burst duration was constant at 15 ms, whereas the burst duration decreased from 15 to 5 ms in one-ms steps for stimuli on the velar-bilabial continuum. We report here examples of results for the synthetic consonant-vowel stimuli, in which listeners were asked to identify the initial consonants.

Figure 3 shows the results for the CV stimuli of the velar-bilabial continuum and the velar-alveolar continuum for nine subjects. Forced-choice categorization functions for each continuum had a reasonably sharp crossover between 100% velar and 100% bilabial or alveolar responses, indicating that most listeners could classify sounds in terms of place of articulation using only the burst spectrum, perhaps relative to certain characteristics of the vowel spectrum. There are some differences between the responses for the velar-alveolar continua with and without aspiration, presumably a consequence of the closer proximity of voicing onset to the burst for the voiced continuum.

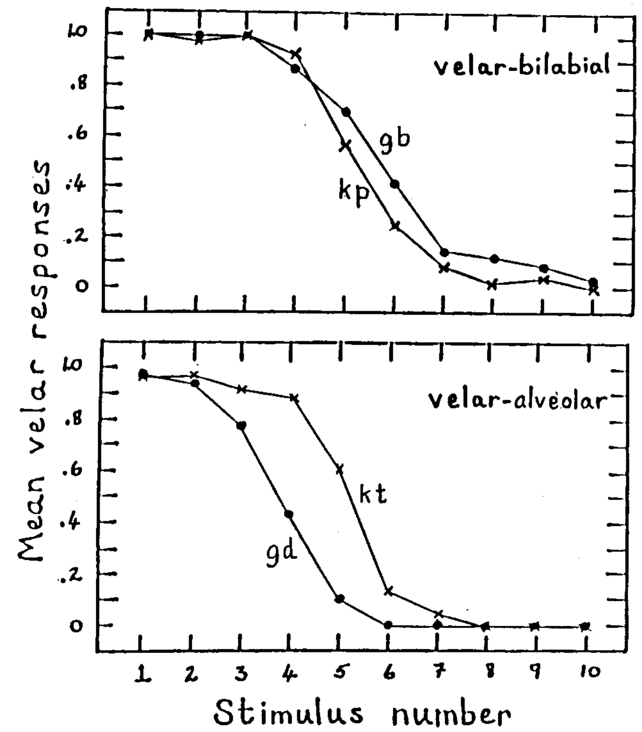


Fig. 3 Responses of nine subjects to stimuli on the velar-bilabial continuum (top panel) and the velar-alveolar continuum (bottom panel). The two functions represent the voiced (filled circles) and voiceless (crosses) continua.

In the upper part of Fig. 4 we show, for the velar-bilabial continuum, the spectrum of the vowel at the vowel onset (light solid line), together with burst spectra for two stimuli: the burst for the most extreme velar (heavy solid line) and the burst for the stimulus at which responses were closest to and not greater than 75% velar (dashed line). A similar display is given at the bottom for the velar-alveolar continuum. In both cases, velar responses for these synthetic stimuli are weakened when the midfrequency prominence drops 3-5 dB below the level of the second formant at the vowel onset. In some sense, the onset of the vowel might function as an anchor, or reference, against which the burst is evaluated.

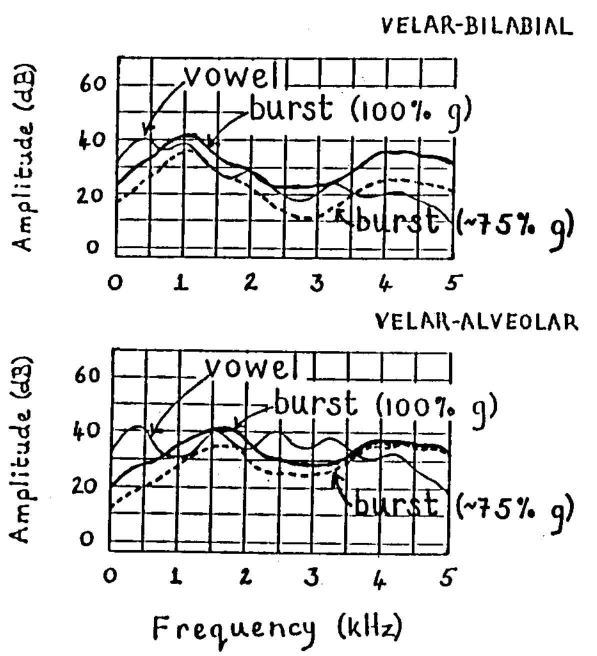


Fig. 4 The three spectra in each panel are the spectrum of the burst at the velar end of the continuum (heavy solid lines), the spectrum of the burst in the stimulus that elicited about 75% velar responses (dashed lines), and the spectrum near the onset of the vowel (light solid lines). The continuum corresponding to each panel is indicated.

Basically this experiment shows that when the midfrequency peak of the burst is sufficiently prominent in relation to the following vowel, listeners identify the consonant as velar even in the absence of transitions. *But* there are some puzzling aspects of these data, and of previous data obtained by others. First, four of our 13 subjects did not hear any velars at all on the velar-alveolar continuum, although their responses to the velar-bilabial continuum were basically the same as those of the other nine subjects. In fact, in presenting the results in Figs. 3 and 4 we have omitted data from the few subjects who heard no velars on the velar-alveolar continuum. These subjects may have focussed on the relation between the spectra of the burst and the vowel at high frequencies, presumably because the midfrequency spectral prominence was not sufficiently salient for them. (In subsequent experiments we increased the amplitude of the midfrequency spectral prominence by about 9 dB at the velar end of the continuum, and all subjects heard velars in this new continuum.) Second, several investigators have shown that velar responses can be obtained when the burst is completely absent, provided that a pair of adjacent formants (usually the second and third formants) are close together at the beginning of the transition into the vowel [3, 4].

These findings, together with published data on the analysis of velar consonants in real speech, have suggested to us that it is an oversimplification to describe a velar consonant as a burst of noise with uniform-amplitude over time, followed by a vowel with suitable transitions. Consequently, we have reexamined the acoustic properties of spoken velar consonants, particularly the fine structure of the short-time spectrum through the burst and into the onset of voicing.

3. Acoustic analysis of natural speech

We have looked at CV syllables spoken by several talkers of British or American English saying /gi/, /ge/, /ga/, /gu/. Several types of spectra were made of the burst and at least the first two periods of the vowel, including Fourier transforms, lpc spectra, and the output of certain auditory models. Spectra were made in successive 5 ms steps and additionally, for some syllables, in smaller steps.

Many of the spectra conform closely to the classical [compact] description for the burst. But as other investigators have also found [2], a substantial minority deviate from the classical picture. However, almost all of these so-called deviant or atypical utterances, have compact properties during at least *some* part of the burst or vocalic onset. Two of the most common types will be shown here.

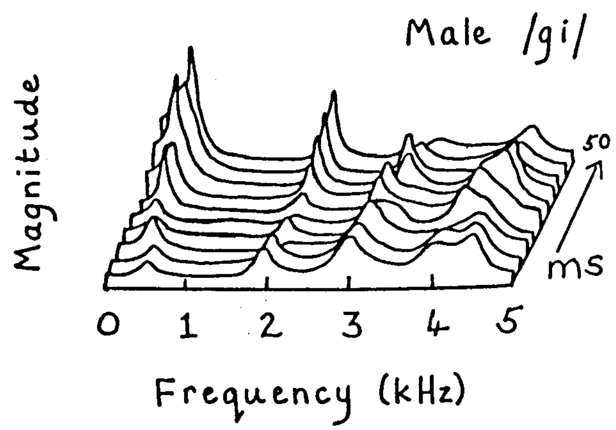


Fig. 5 Sequence of lpc spectra sampled at 5-ms intervals near the release of the syllable /gi/ produced by a male speaker. The amplitude scale is linear. Voicing onset occurs at about the 6th spectrum from the front. This sequence is an example in which spectral prominences appear intermittently in the burst.

Figure 5 shows lpc spectra at successive 5 ms intervals of a male talker's /gi/. Neither the burst nor the vowel onset appear particularly compact, but the peaks and valleys during the burst fluctuate somewhat in amplitude so that a more classical compact spectrum appears intermittently. Such fluctuating spectra will occur when there is a succession of transients at the release, although they can occur even when individual transients cannot be seen in the waveform. These fluctuations in compactness are more common in velars than other stops, presumably because of the longer constriction and slower release, and they may themselves contribute to the perception of velarity.

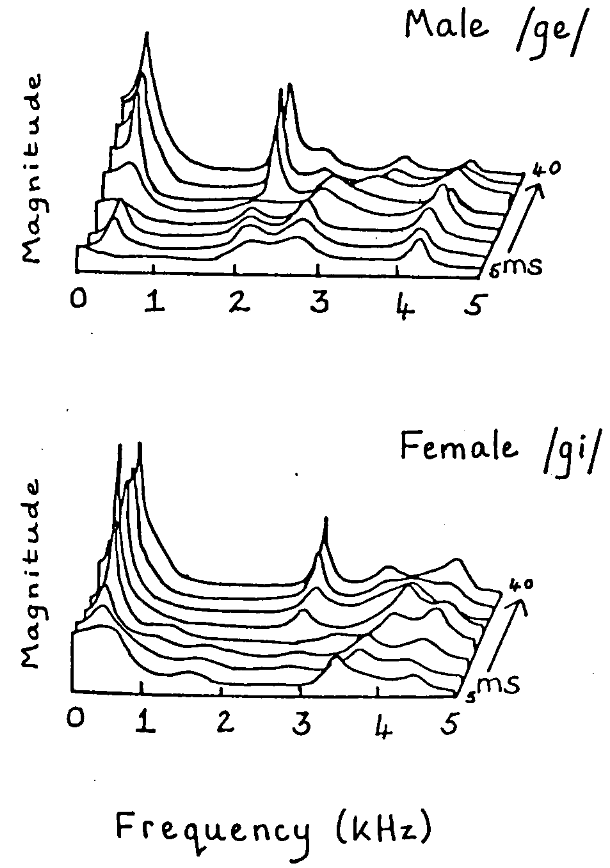


Fig. 6 These are examples of sequences of lpc spectra in which the burst does not show a compact spectral prominence, but a prominence appears following the onset of voicing.

Figure 6 illustrates a second type of nonclassical compactness in which the burst is not compact but there is a very sharp, narrow-bandwidth formant in the midfrequencies at the onset of the *vowel*. This sharply prominent peak appears abruptly and in relative isolation from surrounding peaks, and presumably arises because two formants are very close in frequency. This strongly compact vowel onset appears to be associated more with velars before front vowels, and possibly with weak bursts.

These two phenomena together—a spectrum that fluctuates in degree of compactness during the burst and a strongly compact midfrequency peak at the onset of the vowel—may each serve to enhance the compact percept. Rapid fluctuations in a spectrum may compensate for an otherwise weakly compact burst by somehow focussing attention by virtue of the fluctuating spectrum. And compactneses in the vowel onset may override any ambiguities of the burst. It is worth noting that we often found different types of compactness in different tokens of the same g-vowel syllable, or from the same talker. We are assessing the extent to which these "alternative forms of compactness" can strengthen the perception of velarity in synthetic consonant-vowel syllables.

4. Conclusions

In summary, we have seen in the perception experiments that velar stops in a CV syllable with steady-state formants are heard if the burst has a midfrequency spectral prominence with an amplitude at least as great as that of the corresponding peak at vowel onset. The analyses of natural speech show that the compact prominence is typically present in the burst spectrum, or it may be only intermittent, or it may be more evident in the vowel onset than in the burst. These data suggest that compactness should be defined not in terms of the prominence of a peak in the average burst spectrum, but rather in terms of the occurrence of prominence in the short-time spectrum in at least some region of the syllable onset, whether it be in the burst or in the onset of voicing. One possibility is that the perception of velarity in the consonantal release is enhanced if there are regions in the release phase where a compact spectral prominence is embedded in a context that has reduced compactness or prominence.

If these preliminary observations are confirmed on a larger dataset, then the next task is to begin to describe compactness more precisely through further preception experiments. If we can express compactness in terms of the amplitude, bandwidth, frequency range and time-course of a midfrequency peak relative to adjacent spectra, then we may be on the way to coming up with a description that subsumes burst and transition information under one umbrella.

5. Acknowledgements

This work was supported in part by grants from the National Institute of Neurological and Communicative Disorders and Stroke and from the National Science Foundation.

6. References

1. V.W. Zue. **Acoustic characteristics of stop consonants: A controlled study.** Ph.D. thesis, MIT, Cambridge MA USA (1976).
2. D. Kewley-Port. *Time-varying features as correlates of place of articulation in stop consonants.* **J. Acoust. Soc. Am., 73,** 322-335 (1983).
3. M.F. Dorman, M. Studdert-Kennedy, and L.J. Raphael. *Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues.* **Perception and Psychophysics, 22,** 109-122 (1977).
4. S.E. Blumstein and K.N. Stevens. *Perceptual invariance and onset spectra for stop consonants in different vowel environments.* **J. Acoust. Soc. Am., 67,** 648-662 (1980).

THE PERCEPTION OF VOICING IN DUTCH TWO-OBSTRUENT SEQUENCES

R.J.H. van den Berg

Institute of Phonetics, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

0. Abstract

Perceived voicing in Dutch two-obstruent sequences (C_1C_2), tested in synthetic VCCV nonwords, was shown to depend not only on the amount of periodicity present in the sequence (VOT and VTT), but also on the intensity of frication noise, and on the durations of the second consonant and the preceding vowel. The duration of the first consonant and the speed and range of formant transitions showed no significant effect. Furthermore, these parameters appeared to be independent cues to perception.

1. Introduction

Because of obligatory final devoicing in Dutch no voiced obstruents occur word-finally, e.g. *goed* (good): /xud/ -> /xut/. Therefore, no phonological voicing opposition exists word-finally, and words as *bod* (bid) and *bot* (bone) are phonetically equivalent. As a consequence assimilation in two-obstruent sequences (C_1C_2) with respect to the feature 'voice' can only take place in sequences of which the first consonant (C_1) is voiceless and the second (C_2) is voiced. This initial (i.e. before assimilation takes place) voicing status of the obstruents is the one occurring in an environment in which assimilation cannot take place, e.g. /dIt bukIs xut/ (this book is good).
Assimilation is essentially an articulatory phenomenon [4]. Therefore, if in a sequence of an (initially) voiceless C_1 and voiced C_2 assimilation with respect to 'voice' did take place, both consonants are produced with the same setting of the articulatory feature 'voice', that is both are produced with either vibrating or non-vibrating vocal folds.
Over the years, assimilation of 'voice' in Dutch, which can take place word-internally in compounds as well as across word boundaries, has received a good deal of attention. So far, the aim of the research has been to discover linguistic (and extralinguistic) regularities in the occurrences of assimilation. Two phonological assimilation rules have been formulated.

(1) If C_2 is a plosive, assimilation is regressive, that is C_1 takes on the voicing status of C_2. The result is a sequence of two voiced consonants, e.g. *wit boek* (white book): /wIt buk/ -> /wId buk/.
(2) If C_2 is a fricative, assimilation is progressive, that is C_2 adapts to C_1. The result is a sequence of two voiceless consonants, e.g. *wit zand* (white sand): /wIt zɑnt/ -> /wIt sɑnt/.

These rules were formulated on the basis of data obtained by linguists who listened to utterances, often only to one occurrence as in radio broadcasts or lectures, and noted down cases of assimilation. The decisions about the voicing status of the obstruents were made on the basis of what one heard and they were more often than not made by one perceiver only, the researcher. Moreover, these researchers implicitly assumed that if a voiced (or voiceless) consonant was perceived, a voiced (or voiceless) consonant was produced. However, it is a well-known fact that the perception of voicing is not only affected by the acoustic correlate of presence or absence of vocal fold vibrations, but also by a number of other acoustic cues [6]. In view of this and of the fact that assimilation is an articulatory process, the data obtained by means of this perceptual method can at best be considered as only indirect evidence of assimilation.
A more direct method of establishing whether assimilation did occur would be to measure vocal fold activity during the production of the two-obstruent sequence. Slis [8] took this methodological consequence and performed articulatory/acoustic voice measurements of two-obstruent sequences in which assimilation could occur. The voicing status of the obstruents was established by relating the measurements to those obtained for single voiced and voiceless consonants. In the light of the data thus obtained rule (1) above in particular became contestable.
Slis [9] also made a direct comparison of articulatory/acoustic voice measurements and perceptual voicing judgements of the same natural speech stimuli. From this comparison it appeared that no one-to-one relationship exists between the two types of data: the voicing status assigned on the basis of the presence or absence of vocal fold vibrations was not an adequate predictor of the voicing judgements obtained. However, it is possible that the two consonants did become more alike in some other articulatory feature(s), the acoustic correlates of which may have triggered the perception of two voiced (or two voiceless) consonants.
As stated above, the assimilation rules for Dutch are based on perceptual data. The researchers who formulated them may have been able to distinguish the acoustic correlate of presence or absence of vocal fold vibration from other acoustic cues relevant to the perception of voicing. But these other cues may also have (mis)guided their voicing judgements. Therefore, the study of the relation between acoustic cues and voicing judgements of two-obstruent sequences is of importance for the description of assimilation with respect to the feature 'voice' in Dutch.
The question of what acoustic parameters affect the perception of voicing in two-obstruent sequences was addressed in a series of experiments employing synthetic speech stimuli. Synthetic speech was chosen since it allowed for a ready manipulation of the parameter(s) under investigation and for a strict control on the other parameters. In order not to complicate matters, only one parameter at a time was varied in the earlier experiments, and only after some knowledge was gained about the effects on perception of the various parameters an experiment in which they were covaried was performed. The results of all these experiments are presented below.
Investigated were the effects of voice onset time (VOT), voice termination time (VTT), frication noise intensity, and duration and range of formant transitions [1] as well as duration of C_1 and C_2, and of the preceding vowel [2].

2. Method

The stimuli were generated by a 'speech-synthesis-by-rules' system. In this system a string of phoneme labels and prosodic condition signs is transformed into a string of labels indicating successive segments. Parameter values for each segment are read from a table containing target values and timing data for each parameter (a 'phoneme' representation). These values are adapted for context and prosodic conditions by a set of rules (into an 'allophone' representation). Subsequently, these parameter values for allophone-sized segments are converted into parameter values for segments of pitch period size. These are used as input for the calculation of the synthetic speech signal. The fundamental frequency depended on the intrinsic F_0, stress, and declination, and varied around a mean of about 150 Hz. At the allophone-size level the program allows the user to set parameters at self-chosen values.

2.1. Speech material

To preclude effects due to phonetic context [3] the C_1C_2 sequences were part of VCCV nonwords with a strict control on the vowels. The obstruents included in the research were the labial and alveolar plosives and fricatives. Velar consonants were excluded because in Dutch the phoneme /g/ occurs only in loan words, and the voicing opposition in the velar fricatives is of a doubtful status.
The sequences are described as a voiceless plus voiced consonant, because these were the labels used in the input string for the synthesis system.
Because informal listening showed that synthesizing C_1 plosives with the release burst counteracted the perception of C_1 as voiced, and since I wanted the acoustic signal to be ambiguous with regard to cues that were not under investigation, all C_1 plosives were synthesized without a release burst. For the same reason, the consonantal segments (apart from the stimuli in the VOT and VTT experiments) were synthesized without periodicity but for the first 10 ms of the C_1 segment in which the periodic source amplitude dropped to zero.

2.2. Subjects

In each experiment 12 subjects participated. In the experiment on the durations of C_1 and C_2, and in the covariation experiment the number of subjects was 20. All subjects were university students (ages 19-32) and were paid for their services.

2.3. Procedure

After the stimuli were synthesized they were recorded onto audiotape in random order. Each stimulus was recorded three times in successsion, with a one-second interval between repetitions and an intertrial interval of five seconds, in which the subjects made their response. The subjects' task was to listen to all three repetitions of a stimulus, to identify the consonantal sounds in the sequence, and to indicate in a forced choice task what sequence they had heard. To this purpose the response alternatives were orthographically represented on a score sheet, for example *abda*, *apda*, *apta*, and *abta*, standing for /ɑbdɑ/, /ɑpdɑ/, /ɑptɑ/, and /ɑbtɑ/. The last response category, the voiced-voiceless sequence, which is phonologically inadmissible in Dutch, was not included in the VOT and frication noise intensity experiments. Subjects were tested individually in a sound-treated booth. The stimuli were presented over headphones at a comfortable listening level. Experimental trials were preceded by ten practice trials to allow the subjects to get used to the synthetic speech and to the task. Informal interviews after the tests showed that none of the subjects had experienced difficulties in performing the task, and that all judged the synthetic speech to be of good quality.

2.4. Data analysis

The response categories were labelled (++) for voiced-voiced responses, (-+) for voiceless-voiced, (--) for voiceless-voiceless and (+-) for voiced-voiceless responses. For each stimulus the frequencies of the response categories were assessed. The resulting matrix of frequencies was analyzed according to Goodman's loglinear model [5]. This model was specifically developed for frequency data with more than one independent variable. The statistic employed is a multidimensional chi square.

3. Voice Onset Time (VOT)

The effect of VOT was tested in 16 sequences, all combinations of /p,t,s,f/ + /b,d,z,v/. A uniform VOT continuum was opted for, and therefore the durations of the consonantal segments were set at a constant value of 75 ms each, counting the silent interval of the plosives and the noise portion of the fricatives. The durations of the preceding and following vowel were set at 90 and 170 ms respectively. Five VOT values were employed: -150, -75, -30, 0, and +20 ms. These values are relative to the end of the segment representing C_2, that is the end of the silent interval for plosives and the moment of frication noise offset for fricatives.
The effect of VOT was significant: $\chi^2= 246.23$, df=8, $p<0.001$. VOT did not interact with sequence type: the same general pattern was observed for all sequences. From the data in Table 1 it is clear that late VOT's (that is with no periodicity present in the consonantal segments) favour (--) responses, as could be expected. With earlier VOT's the responses shift via (-+) towards (++).

Table 1: Response frequencies for five VOT's (in %).

VOT	(++)	(-+)	(--)
-150	55.7	33.9	10.4
-75	39.6	46.4	14.1
-30	21.4	52.6	26.0
0	13.5	29.7	56.8
+20	9.9	26.6	63.5

Comparison of the -30 and 0 ms VOT conditions shows that the 30 ms stretch of periodicity at the end of the C_2 segment is a strong cue to C_2 perception: it raises the number of [+voice] C_2 percepts by 30.8%. But the number of [+voice] C_1 percepts also increases, although to a lesser extent (7.9%). Apparently, the stretch of periodicity is also taken as a voicing cue to C_1. This seems to indicate that cues from rather distant portions of the acoustic signal can be integrated into a perceptual unit. However, the effect of this cue appears to become reduced with increasing distance.

4. Voice Termination Time (VTT)

In testing the effect of VTT the same 16 sequences were used. The durations of the segments were as in the VOT experiment. Five VTT values were employed:

0, 40, 75, 110, and 150 ms. These values are relative to the beginning of the C_1 segment.
As can be seen from Table 2 VTT affects perception significantly: χ^2= 451.16, df=8, p<0.001. Again there was no interaction with sequence type: the same pattern was found for all sequences. The responses were in line with what was expected: short VTT's led to (--) and long VTT's to (++) responses, whereas for the intermediate VTT's (40 and 75 ms) the highest frequencies for (+-) were observed, although the frequencies for (++) and (-+) also increased in comparison with a VTT of 0 ms. This is most likely due to the fact that a voiced-voiceless sequence is phonologically inadmissible in Dutch. Probably, phonological restrictions affected perception [7], and the subjects (having perceived periodicity) resorted to the (++) and (-+) categories.

Table 2: Response frequencies for five VTT's (in %).

VTT	(++)	(-+)	(--)	(+-)
0	3.1	16.7	66.1	14.1
40	6.8	24.5	37.5	31.3
75	28.6	32.3	14.1	25.0
110	55.7	32.3	4.2	7.8
150	62.5	31.8	2.6	3.1

These data, too, show indications that cues from rather distant parts of the acoustic signal are integrated into a perceptual unit. The 40 ms stretch of periodicity at the beginning of the C_1 segment not only raised the number of [+voice] C_1 percepts by 20.9%, but also the number of [+voice] C_2 percepts by 11.5%.

5. Frication Noise Intensity

Again the same 16 sequences were tested. The durations of all segments were controlled by the timing rules of the synthesis system. Since C_1 plosives contained no release burst, the variation of the noise intensity was restricted to C_2 (the noise portion of the fricatives and the release burst of the plosives). The six amplitude values used were chosen so as to cover the voiced-voiceless continuum without exceeding naturalness limits. This resulted in a 3 dB step size for fricatives and 6 dB steps for plosives.

Table 3: Response frequencies for six noise level values (in %). Obstruent-plosives only.

noise	(++)	(-+)	(--)
low	21.9	35.4	42.7
	19.8	36.5	43.8
|	11.5	38.5	50.0
V	6.3	36.5	57.3
	12.5	22.9	64.6
high	6.3	27.1	66.7

The overall effect of noise intensity was significant: χ^2= 30.74, df=10, p<0.01. However, the interaction with sequence type was also significant: noise intensity did show an effect for obstruent-plosives, but not for obstruent-fricatives. The difference is most likely due to the different step sizes involved. The direction of the effect is as was expected: with increasingly higher noise levels more (--) responses were given at the cost of (++) and (-+). As may be clear from the data in Table 3, the effect was rather weak, which may have been due to the fact that other parameters were not set at adequate values. Most likely the segmental durations, controlled by the built-in synthesis rules, were too long and biased the responses to (--) and (-+).

6. Formant transitions

Range and duration of the F1, F2, and F3 transitions into and out of the C_1C_2 sequence were tested in /pd/, /tb/, /fd/, and /sb/. The consonantal durations were set at 60+65 ms for /pd/ and /tb/, and at 70+70 ms for /fd/ and /sb/. Speed and range of the transitions were controlled by setting the moment of transition onset and the time within which the shift takes place. For VC_1 transitions moment of onset was relative to the beginning of the C_1 segment, for C_2V transitions it was relative to the boundary between the silent interval and the release burst.
Only a limited range of values could be used, because in informal listening it appeared that long transitions led to the perception of glides and short transitions to the loss of the principal perceptual cue to the place of articulation. Three types of VC_1 transitions: -40;40, -40;60, and -20;40 (moment of onset and transition time respectively) were combined with four types of C_2V transitions: -10;95, -10;75, 0;75, and 0;55.
No overall effect on perception was observed: χ^2= 31.10, df=33, ns. Also, the interaction between VC_1 and C_2V transitions and the main effect of C_2V transitions were not significant. A small effect of VC_1 transitions (p<0.05) appeared to be due to one sequence only (/pd/), the responses to which showed no coherent pattern.

7. Durations of C_1 and C_2

Since the durations of both C_1 and C_2 were varied it seemed advisable to have a clear acoustic boundary between the two consonants. So, only combinations of a fricative (noise) with a plosive (silent interval) were used: /fd,sb,pz,tv/. The durations of the preceding and following vowel were set at 90 and 160 ms respectively. While limiting the total duration of the sequence to 150 ms, both the durations of C_1 and C_2 were varied in 5 steps of 15 ms starting at 45 ms. This resulted in 15 combinations of durations (C_1 and C_2 respectively): 45-45, 45-60, ... 45-105, 60-45, ... 60-90, ... 105-45. All stimuli were synthesized under two stress conditions: stress on the first or on the second syllable. Stress was synthesized by means of a prominence lending rise and fall of F_0.
The interactions of sequence type and stress pattern with duration were not significant. This signifies that duration had a similar effect for the various sequences and both stress patterns, so for the duration results the data were pooled over these conditions. Since the design was not fully crossed for C_1 and C_2 duration, the effect of C_1 duration could only be tested for the various levels of C_2 duration separately, and the effect of C_2 duration only for the various levels of C_1 duration separately.
For none of the levels of the C_2 duration variable did the factor of C_1 duration affect the frequency distribution of the four response categories. Even if only the responses to C_1 were considered, no significant effect was obtained. C_2 responses, too, were not affected by C_1 duration.
On the other hand, for a C_1 of 45 ms (C_2 ranging from 45 to 105 ms), for a C_1 of 60 ms (C_2: 45-90 ms), and for a C_1 of 75 ms (C_2: 45-75 ms) the effect of C_2 duration was significant. Longer C_2 durations led to more (--) and (+-), and less (++) and (-+) responses. The picture becomes even clearer if C_2 responses only are considered: longer C_2 durations led to more [-voice] C_2 percepts. No effect of C_1 duration on C_1 perception was observed.
So, what the effect of this manipulation seems to be boiling down to is that C_2 duration affects C_2 perception, longer durations giving rise to more [-voice] percepts. This effect is strong enough to affect the frequency distribution of the four response categories, wheras C_1 duration does not have any effect at all.
The significant effect of stress pattern manifested itself in that more (++) and less (--) responses were given if stress was on the first than if it was on the second syllable.

8. Preceding Vowel Duration

This effect was tested in the sequences /pd, tb, fd, sb, pz, tv, fz, sv/. The vowels in the VCCV nonwords were either a phonologically long /a:/ or a phonologically short /ε/, to test a possibly differential effect of preceding vowel duration for vowels of different phonological length. The durations of /pd/ and /tb/ were 60+65 ms, of the other sequences 70+70 ms. The following vowel had a duration of 160 ms. Stress was either on the first or on the second syllable (stress-1 and stress-2 respectively), realized by a prominence lending rise and fall. To avoid a clash between stress and duration longer preceding vowel durations were used in the stress-1 condition (80, 120, and 180 ms), than in the stress-2 condition (55, 80, and 120 ms).
Vowel type and sequence type did not interact with vowel duration, so the data were pooled over these conditions. Preceding vowel duration significantly (χ^2= 21.25, df=6, p<0.001) affected the response distribution in the stress-1 condition, but not in the stress-2 condition. Under stress-1 longer durations led to more (++) and (+-), and to less (-+) and (--) responses (see Table 4).

Table 4: Response frequencies for three preceding vowel durations (in %).

stress-1	(++)	(-+)	(--)	(+-)	C_1=(+)	C_2=(+)
80	10.4	31.8	38.0	19.8	30.2	42.2
120	17.7	30.2	31.3	20.8	38.5	47.9
180	26.6	22.4	27.6	23.4	50.0	49.0
stress-2	**(++)**	**(-+)**	**(--)**	**(+-)**	**C_1=(+)**	**C_2=(+)**
55	12.5	27.6	45.3	14.6	27.1	40.1
80	10.4	28.6	42.2	18.8	29.2	39.1
120	15.6	29.2	31.3	24.0	39.6	44.8

From a comparison of the 80 and 120 ms conditions under stress-1 with the same conditions under stress-2, it appeared that the non-significant effect for stress-2 is not due to the difference in stress, but rather to the smaller absolute range in durations. However, if only C_1 responses are considered, the effect of vowel duration is significant for stress-1 (χ^2= 15.82, df=2, p<0.001) as well as for stress-2 (χ^2= 7.76, df=2, p<0.05), although the effect is still larger for stress-1. Perception of C_2 was not affected significantly in both stress conditions.
So, preceding vowel duration affects C_1 perception, with longer durations leading to more [+voice] C_1 percepts. In stress-1 this effect is large enough to influence the distribution of the four response categories.
The interaction vowel type x duration was not significant. However, since vowel duration mainly affected C_1 responses, the interaction was also tested with C_1 responses as the dependent variable. For stress-1 the interaction was significant: χ^2= 6.78, df=2, p<0.05. With increasing vowel duration the number of [+voice] percepts increased more rapidly for /ε/ than for /a:/, so it seems that an /ε/ is perceived as longer than an /a:/ of the same duration. From this it may be inferred that an internal representation of a vowel's intrinsic duration might play a role in its perceived duration.

9. Covariation of parameters

To study possible interactions between parameters an experiment was run in which those parameters that showed a significant effect were covaried. Four C_1C_2 sequences were used: /fd,sb,pz,tv/. The duration of the C_1 segment was 50 ms, that of the second vowel 160 ms. The six parameters that were varied were VOT (in three steps of 0, -25, and -50 ms), VTT (in two steps of 0 and 40 ms), noise intensity of the fricatives (two levels with a difference of approximately 10 dB), duration of C_2 (in three steps of 50, 75, and 100 ms), duration of the preceding vowel (in two steps of 80 and 120 ms), and stress pattern (stress on the first or on the second syllable).
Some parameters interacted with sequence type due to the fact that for some sequences the effect of some parameters was more powerful, rather than to a totally different response pattern. Therefore, the data were analyzed for all four sequences separately. It appeared that for each sequence all six factors had a highly significant effect (p<0.001) on perception and that the response patterns were in line with the earlier findings. The few significant interactions that were obtained seemed to be incidental, since, if an interaction was observed, it was found for one out of four sequences only.

10. Conclusion

These results show that perception of voicing in Dutch two-obstruent sequences does not depend solely on the presence/absence of periodicity. Furthermore, it seems justified to conclude that the factors that affect the perception of voicing in two-obstruent sequences (viz. VOT, VTT, frication noise intensity, stress pattern, C_2 duration, and preceding vowel duration) do so independently.

References

[1] R.van den Berg, "The effect of varying voice and noise parameters on the perception of voicing in Dutch two-obstruent sequences", *Speech Communication 5(4)*, 355-367, 1986.
[2] R.van den Berg, "Effects of duration on the perception of voicing in Dutch two-obstruent sequences", *J.Phonetics*, subm.
[3] R.van den Berg, I.Slis, "Phonetic context effects in the perception of voicing in C_1C_2 sequences", *J.Phonetics 15(1)*, 39-46, 1987.
[4] A.Crystal, "*A first dictionary of linguistics and phonetics*", Andre Deutsch, London, 1980.
[5] L.Goodman, "The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach", *Biometrika 60*: 179-192, 1973.
[6] L.Lisker, "Rapid vs Rabid: a catalogue of acoustic features that may cue the distinction", *Haskins Lab.Stat.Rep.Speech Res. SR-54*: 127-132, 1978.
[7] D.Massaro, M.Cohen, "Phonological context in speech perception", *Perception and Psychophysics 34(4)*: 338-348, 1983.
[8] I.Slis, "Assimilatie van stem in het Nederlands", *Glot 5*: 235-261, 1982. Also as: I.Slis, "Rules for assimilation of voice in Dutch", in: R.Channon & L.Shockey (eds) "*In honour of Ilse Lehiste/Ilse Lehiste Pühentusteots*", Foris Publications, Dordrecht, 225-240, 1986.
[9] I.Slis, "Assimilation of voice in Dutch as a function of stress, word boundaries, and sex of speaker and listener", *J.Phonetics 14*: 311-326, 1986.

ASSIMILATION OF VOICE AND PERCEPTION OF VOICING: EFFECTS OF PHONETIC CONTEXT

I.H. Slis & R.J.H. van den Berg

Institute of Phonetics, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

0. Abstract

The results of an earlier experiment contained indications that the degree of voicing in the phonetic context affected the perception of voicing in Dutch two-obstruent sequences. This was confirmed in a separate perception experiment. The articulatory/acoustic measurements obtained in a production experiment refute an explanation in terms of a perception mechanism in which regularities in speech production are embodied. The phonetic context effect appears to be a purely perceptual phenomenon.

1. Introduction

Up to only a few decades ago, assimilation of voice in Dutch two-obstruent sequences was investigated by linguists who scored instances of assimilation by ear, often after a single presentation of the utterance. The results of these investigations varied considerably, leading to a great variation of opinions upon the subject [12]. In this contribution we will try to show that one of the possible causes for this lack of agreement may be found in the phonetic context of the two-obstruent sequences.
According to Crystal's [7] definition, assimilation is '*the influence of one sound segment upon the articulation of another so that the two sounds become more alike, or identical*'. In line with this definition, we too consider assimilation to be an essentially articulatory phenomenon. If in a two-obstruent sequence assimilation of voice takes place, both consonants will be produced with the same vocal fold setting: vibrating or non-vibrating. This point of view was the basis for a number of articulatory/acoustic measurements relating to vocal fold behaviour during the production of the obstruent sequence [12].
The consonantal sequence in which assimilation of voice has taken place may be perceived as a sequence of two voiced (or two voiceless) consonants. However, one may perceive two consonants as having the same voicing status in spite of the fact that assimilation of voice did not take place [14]. In that case it is obvious that voicing cues (i.e. acoustic cues to the voicing status of the consonants in question) other than the auditory result of the presence or absence of vocal fold vibration are used by the listener. At the Institute of Phonetics in Nijmegen (IFN) the effect of voicing cues on the perception of voicing in two-obstruent sequences is being investigated in a series of experiments. As was the case with single voiced and voiceless obstruents in Dutch [13], a number of cues were found to affect the perception of voicing in such C_1C_2 sequences [1,2,3].
One of the factors that may affect the perception of voicing in C_1C_2 sequences appears to be the degree of voicing in the consonants in the context. Indications to that effect were found in an earlier study [4] set up to investigate the most suitable type of stimuli for a large series of experiments on the perception of voicing in C_1C_2 sequences. In this paper we will briefly discuss this study (section 2). The results gave rise to an experiment specifically designed to investigate such phonetic context effects on the perception of voicing in C_1C_2 sequences [5]. This experiment is presented in section 3 of this paper. In section 4, several hypotheses will be forwarded that may explain the results obtained. In order to be able to choose between the hypotheses a production experiment was run, which is discussed in section 5.
All experiments employed heterorganic two-obstruent sequences (C_1C_2) to avoid problems arising from the use of (homorganic) geminates. Because of restrictions inherent in Dutch [6] the sequences consisted of a phonologically voiceless obstruent (C_1) followed by a phonologically voiced one (C_2).

2. Investigation of optimal stimulus form

In this first experiment [4] we investigated the perception of voicing in two-obstruent sequences that were part of two successive syllables (C_iVC_1-C_2VC_f). One of the aims of this study was to investigate whether the linguistic status of the stimuli would affect the perception of voicing in such sequences. To this purpose the C_1C_2 sequences were embedded in three types of linguistic context:
(a) a word pair that was part of a meaningful sentence;
(b) the same word pair in isolation;
(c) an utterance made up of two meaningless syllables; these nonwords were obtained by changing the initial consonant (C_i) of the first word and the final consonant (C_f) of the second word of the same pairs as used in conditions (a) and (b).

All stimuli were generated by means of a speech-synthesis-by-rules system available at the Institute of Phonetics Nijmegen [11]. Eighteen subjects participated, who identified the consonants and indicated what sequence they had heard in a forced choice task with four response alternatives: voiced-voiced, notation (++); voiceless-voiced, notation (-+); voiceless-voiceless, notation (--); and voiced-voiceless, notation (+-). This last sequence is irregular in Dutch according to the generally accepted phonological rules, but it was nevertheless included, because the subjects felt the need for this response category.

No differences in the perception of voicing in C_1C_2 sequences were observed between the sentence and word pair conditions. However, a significant (χ^2= 12.30, df=6, p<.01) difference was found between word pairs and nonwords.
Three possible explanations for this difference offer themselves:

1) A lexical explanation: the listener is inclined to interpret the perceived sounds so that they make up an existing word. We may expect, therefore, that the responses show a bias towards the perception of meaningful words, and consequently towards the perception of a voiceless consonant followed by a voiced one, which yields a string of unaltered words. In those cases where a nonword can be changed into a word by a shift in the voicing status of one of the members of the C_1C_2 sequence, the meaningful word is expected to prevail. A shift away from the voiceless-voiced responses can then be expected. Nothing of the kind is observed; on the contrary, nonwords show more voiceless-voiced responses than word pairs.
2) A phonological explanation: the listener's perception is subject to his knowledge of phonological rules, particularly in a 'language-mode' of listening. Therefore, we expect that the subjects will perceive more 'regular' sequences in word pairs than in nonwords. So, we expect a higher number of voiced-voiced responses in word pairs with obstruent-stop sequences, and a higher number of voiceless-voiceless responses in word pairs with obstruent-fricative sequences [6]. The results are to the contrary: we observed more irregular sequences (voiced-voiceless) in words than in nonwords.
3) A phonetic explanation: a change in the sound structure of the context might have affected perception. Since the linguistic and phonological explanations did not adequately predict the observed response patterns, we were left with the phonetic explanation. The nonwords were derived from the word pairs by altering the initial (C_i) and the final (C_f) consonant. Therefore, the only phonetic difference between the word pairs and the nonwords was in the C_i and the C_f. So, if a difference in the phonetic context affects the perception of voicing in the C_1C_2 sequences, it is obvious that the alterations in C_i and C_f must be the cause for the perceptual differences observed.

An analysis of the results showed that in those cases where C_i and/or C_f was changed into a voiceless consonant, the number of responses containing voiceless C_1's and C_2's increased. A more detailed analysis suggested that changes in the voicing status of C_i were related to changes in C_1 responses, and changes in the voicing status of C_f to changes in C_2 responses. Since this phonetic context effect was not expected, we had not controlled for the voicing status of the phonetic context when generating the stimulus material. In order to investigate the effect more systematically, a new experiment, specifically designed for this purpose, was carried out.

3. The effect of voiced/voiceless contexts on the perception of voicing in C_1C_2 sequences

The effect of voicing in the phonetic context on the perception of voicing in two-obstruent sequences was investigated in synthetically generated nonwords of the type $C_iVC_1C_2VC_f$. Both syllables were stressed, the first by a rise, the second by a fall in the fundamental frequency contour [8]. The vowel in both syllables was an /ɑ/. The phonetic context, the independent variable, was formed by C_i and C_f. Both could be either an /s/ (voiceless context), or an /n/ (voiced context). The C_1C_2 sequences used were all possible heterorganic combinations of labial and dental obstruents, viz. /pd, tb, fd, sb, pz, tv, fz, sv/. On the basis of the results of previous experiments the synthesis parameters were chosen so as to yield stimuli that were ambiguous with respect to the perceptual voicing status of C_1 and C_2. This implies that the stop-stop sequences had a closure interval of 125 ms, and the other sequences one of 140 ms. The stimuli were synthesized without periodicity during the closure interval. Procedure and response categories were as described above.

Table 1: Frequencies of perceived voicing in C_1C_2 sequences as a function of the voicing status of the context (in %).

context	(++)	(-+)	(--)	(+-)
n--n	33.6	44.8	10.3	11.4
s--n	19.5	61.6	12.2	6.7
s--s	14.5	50.3	27.2	8.0
n--s	26.3	35.2	22.0	16.6

The results (see Tables 1 and 2) showed a highly significant effect of voicing status of the phonetic context. With a voiced C_i, viz. /n/, a significantly (χ^2=107.77, df=1, p<.001) higher number of voiced C_1 percepts was observed than with a voiceless C_i, viz. /s/. With a voiced C_f (/n/) significantly (χ^2=86.87, df=1, p<.001) more voiced C_2's were perceived than with a voiceless C_f (/s/).
The voicing status of C_i was found to have no significant effect on the perception of C_2, nor did the voicing status of C_f affect C_1 perception. Therefore, it would seem that effects of voicing in the context are restricted to the syllable. However, it remains possible that such effects can occur over longer temporal distances, and thus across syllable boundaries.

Table 2: Frequencies of C_1 and C_2 responses as a function of the voicing status of the initial and final context (in %).

context	C_1=(+)	C_1=(-)	C_2=(+)	C_2=(-)
n--	43.8	56.2	69.9	30.1
s--	24.4	75.6	73.0	27.0
--n	35.5	64.5	79.8	20.2
--s	32.7	67.3	63.1	36.9

4. Discussion

In this section we will discuss four different hypotheses that may explain the results obtained. The first two are based on the assumption that the perceptual mechanism uses its awareness of regularities in speech production. The other two hypotheses are purely perceptual in nature.

A1) Perceptual compensation of coarticulatory differences

Let us assume that a difference in the degree of voicing in the context leads to a different production of the C_1C_2 sequence. In that case, there is a ground for a mechanism like perceptual compensation. According to this mechanism listeners perform perceptual corrections for differences in the production of natural speech that arise from contextual influences [10]. The result of these corrections is that no differences are perceived. In synthetic speech stimuli in which these articulatory/acoustic differences are absent, the same compensation mechanism will lead to perceptual differences. When we apply this to voicing in the context, we come to the following argument.
In order to explain our present results, we would have to assume that in natural speech C_1 (or C_2) is

produced with stronger 'voicing' if C_i (or C_f) is voiceless as compared to the condition where C_i (or C_f) is voiced. This may be seen as a kind of emphasized articulatory contrast. The listener's compensation for this articulatory contrast will lead to the perception of the same sequence in all contexts. If, however, the stimulus is ambiguous, as was the case in the present experiment, the same mechanism will result in the perception of a more voiceless sequence in a voiceless context, and a more voiced sequence in a voiced context.

A2) Perceptual expectation of coarticulatory effects

The listener may be inclined to perceive the things he expects, in other words he may be the victim of selective perception. In order to explain our results we must assume that the listener expects to hear a voiceless C_1 (or C_2) in combination with a voiceless C_i (or C_f). This expectation must be based on facts in natural speech. Therefore, we have to assume that voiceless consonants in the context lead to devoicing of (some of) the nearby consonants. From a coarticulatory viewpoint this is a plausible position.

The hypotheses A1 and A2 are mutually exclusive, since they assume opposite effects in production. So, articulatory measurements must enable us to make a choice between the two, or, in case no differences are found, refute them both.

B1) A perceptual-phonological explanation

In this purely perceptual hypothesis we assume that a sequence of speech sounds is recognized in terms of a sequence of 'bundles of phonological features' to which phoneme labels are attached. Context effects as the one found in Experiment 2 may occur when a correctly identified feature of C_i or (C_f) is erroneously attributed to C_1 or (C_2). If this type of erroneous attributions in fact occur at the phonological level, it is likely that the acoustic duration of the intervening phoneme (the vowel) is of no consequence. In that case no effect of (intervening) vowel length is expected. A second factor that can be expected to induce this type of attribution errors is the resemblance between the two 'phonological feature bundles'. In a way similar to the processes involved in producing slips of the tongue, we may expect an increase in the number of attribution errors if the two phonemes (context and target) have more features in common.

B2) A perceptual-phonetic explanation

In this hypothesis we assume that the error occurs at a more peripheral level, viz. that of acoustic cue integration. During this stage the acoustic cues are held in a preperceptual auditory storage (PAS) [9]. The time span of PAS is about 200-250 ms. So, if the temporal distance between context phoneme and target phoneme is less than this time span, the cues for the two phonemes are simultaneously present in PAS. In such a situation misattributions of cues may occur, resulting in a cue being erroneously taken as a voicing cue to the wrong sound segment. In this way effects of voicing in the phonetic context on the perception of voicing in C_1C_2 sequences may be explained. Assuming that the 'strength' of cues plays a role, we expect that such errors are likely to be more frequent with an increase in 'cue strength' (voicedness or voicelessness) of the context phonemes. So, the frequency of erroneous cue attribution may be expected to be dependent on the degree of voicedness or voicelessness. The notion 'degree of voicing' may, for example, be operationalized as the position of C_i (or C_f) on a 'voicing scale' depending on e.g. VOT (or VTT). Furthermore, a greater temporal proximity of the context phoneme and target phoneme may also promote misattributions of voicing cues. So, the (phonological) duration of the vowels intervening the context phoneme and the target phoneme, that is the V's in a $C_iVC_1C_2VC_f$ sequence, is expected to interact with the effect of voicing in the phonetic context on the perception of voicing in C_1C_2 sequences.

The hypotheses B1 and B2 make different predictions with respect to the effect of vowel duration and with respect to a possible effect of the position of C_i (or C_f) on the voicing scale.

In order to decide which of the four hypotheses outlined above prevails, two more experiments need to be carried out. The first one, a production experiment, will enable us to find out whether the context affects articulation of the C_1C_2 sequence. If articulatory effects are found, we may decide on hypothesis A1 (perceptual compensation) if a voiceless context (C_i and/or C_f) leads to more voicing in C_1C_2, or on hypothesis A2 (perceptual expectation) if it leads to less voicing in C_1C_2.
If no context effects are found in articulation there is no ground to maintain hypotheses A1 and A2, and a purely perceptual explanation would seem appropriate. The crucial experiment for the choice between hypotheses B1 and B2 would be one in which the degree of voicing in the context and the duration of the intervening vowels are varied. If perceptual errors result from an erroneous attribution of already recognized phonological features, neither gradations of voicing in the context, nor the durations of the vowel phonemes are expected to affect the perception of voicing in the C_1C_2 sequence. If, on the other hand, the errors are located in the cue integration stage, we expect to find effects of gradation of voicing and of vowel length.
The production experiment will be discussed in the next section. The perception experiment has as yet not been carried out.

5. The effect of voicing in the context on the production of C_1C_2 sequences

The production experiment did in fact consist of three parts, referred to as part (a), (b), and (c), respectively. In each of the three, five male speakers participated, who were asked to read the stimulus materials. The acoustic signal was recorded via a microphone, vocal fold activity by means of an electrolaryngograph. Both these signals were registered on photographic paper with a UV-recorder (SE oscillograph 6008). In the oscillograms we related the moment of voice termination (VTT) to that of oral closure, and the moment of voice onset (VOT) to that of oral release. According to criteria derived from single voiced and voiceless consonants, the voicing status of C_1 and C_2 was assessed, and thus whether assimilation of voice had occurred or not. For a detailed description of the procedure and criteria, see [12].

In part (a) the stimuli were the same as in the previous experiment, embedded in a short carrier phrase, viz. 'doe die $C_iVC_1C_2VC_f$ om'. Employing this type of stimuli resulted in a very low frequency of produced assimilation in obstruent-stop sequences (all obstruent-fricative sequences were progressively assimilated). This was probably due to the fact that the speakers were aware of the central role of the nonword (the only element to vary in the sentences) and may therefore have been inclined to pronounce it with great care. So, we had five other

speakers read two additional series (parts (b) and (c)) consisting of meaningful sentences in which the C_1C_2 sequence was part of two adjacent words (C_iVC_1-C_2VC_f). In these sentences C_i and C_f were either voiceless (single consonants or consonant clusters) or voiced (single nasals). The C_1C_2 sequences used in these two series were all heterorganic obstruent-stop sequences. In part (b) the C_1C_2 sequence followed a stressed syllable, in part (c) it preceded a stressed syllable. On the text sheet syllables that had to be stressed were underlined. The speakers were instructed to read the sentences as spontaneously as possible.

Table 3: Frequencies of assimilation of voice as a function of voicing in the context (in %).

context	(++)	(-+)	(--)
+...+	31.7	41.7	26.7
-...+	30.0	50.0	20.0
-...-	31.7	38.3	30.0
+...-	26.7	33.3	40.0

In contrast with part (a) assimilation of voice, either regressively (i.e. two voiced consonants) or progressively (i.e. two voiceless consonants) occurred rather frequently. In line with earlier measurements [14] stress on the syllable preceding the C_1C_2 sequence (part (b)) favoured progressive assimilation, and stress on the following syllable (part (c)) favoured regressive assimilation. However, in none of the three parts of this experiment did we observe a significant effect of voicing in the context on the production of the C_1C_2 sequences, that is on assimilation of voice in those sequences. For this reason, and because the number of speakers was rather low, we pooled the obstruent-stop data from the three parts of the experiment. These pooled data are given in Tables 3 and 4. As may be clear from the figures no significant context effect was found.

Table 4: Frequencies of produced voiced and voiceless C_1 and C_2 as a function of voicing in initial and final context (in %).

context	C_1=(+)	C_1=(-)	C_2=(+)	C_2=(-)
+...	29.2	70.8	66.7	33.3
-...	30.8	69.2	75.0	25.0
...+	30.8	69.2	76.7	23.3
...-	29.2	70.8	65.0	35.0

6. Conclusion

Since we did not find any effects of voicing in the phonetic context (C_i and/or C_f) on the production of the two-consonant sequence C_1C_2, we conclude that these results refute the first two hypotheses, viz. perceptual compensation (A1) and perceptual expectation (A2) of articulatory differences. Thus we are left with the two purely perceptual hypotheses (B1 and B2). The question of whether we have to look for an explanation in terms of an erroneous attribution of a phonological feature (that is to the wrong phoneme), or whether the error occurs at the cue integration stage, cannot be settled by the present data. To address this issue an experiment needs to be run in which the degree of voicedness/voicelessness in the context (C_i and/or C_f) is systematically varied by choosing C_i and C_f from a continuum. Besides, by varying the time interval between context and target phoneme, we may be able to assess whether the domain over which phonetic context effects do take place is determined in durational terms or in terms of number of phonemes, and thus whether the effect originates in PAS, or from misattributions on a higher, phonological level.

References

[1] R.van den Berg, "The effect of varying voice and noise parameters on the perception of voicing in Dutch two-obstruent sequences", *Speech Communication 5(4)*, 355-367, 1986.
[2] R.van den Berg, "Effects of duration on the perception of voicing in Dutch two-obstruent sequences", *J.Phonetics*, subm.
[3] R.van den Berg, "The perception of voicing in Dutch two-obstruent sequences", *this issue*, 1987.
[4] R.van den Berg, I.Slis, "Perception of assimilation of voice as a function of segmental duration and linguistic context", *Phonetica 42(1)*, 25-38, 1985.
[5] R.van den Berg, I.Slis, "Phonetic context effects in the perception of voicing in C_1C_2 sequences", *J.Phonetics 15(1)*, 39-46, 1987.
[6] G.Booij, "*Generatieve fonologie van het Nederlands*", Het Spectrum, Utrecht/Antwerpen, 1981.
[7] A.Crystal, "*A first dictionary of linguistics and phonetics*", Andre Deutsch, London, 1980.
[8] A.van Katwijk, "*Accentuation in Dutch: an experimental linguistic study*", van Gorcum, Amsterdam/Assen, 1974.
[9] D.Massaro, "*Experimental Psychology and Information Processes*", Rand McNally, Chicago, 1975.
[10] B.Repp, "Perceptual integration and differentiation of spectral cues for intervocalic stop consonants", *Perception and Psychophysics 24(5)*, 471-485, 1978.
[11] I.Slis, "Some remarks on speech synthesis by rule", *Proceedings Institute of Phonetics, University of Nijmegen 2*, 83-99, 1978.
[12] I.Slis, "Assimilatie van stem in het Nederlands", *Glot 5*: 235-261, 1982. Also as: I.Slis, "Rules for assimilation of voice in Dutch", in: R.Channon and L.Shockey (eds.) *In honour of Ilse Lehiste/Ilse Lehiste Pühentusteots*, Foris Publications, Dordrecht/Cinnaminson, 225-240, 1986.
[13] I.Slis, "*The voiced-voiceless distinction and assimilation of voice in Dutch*", Unpublished Doctoral Dissertation, University of Nijmegen. SR-54: 127-132, 1985.
[14] I.Slis, "Assimilation of voice in Dutch as a function of stress, word boundaries, and sex of speaker and listener", *J.Phonetics 14*: 311-326, 1986.

CUE-TRADING RELATIONS FOR INITIAL STOP VOICING CONTRAST AT DIFFERENT LINGUISTIC LEVELS

U THEIN-TUN

School of Communication Disorders
Lincoln Institute
Melbourne, Australia 3053

ABSTRACT

The speech mode of processing is a special mode of information processing in the sense that the cue-trading relationship of multiple cues signifying one phonemic contrast is most effective in the speech mode but not in the sense that cue-trading does not exist at non-linguistic levels.

INTRODUCTION

Fitch, Halwes, Erickson and Liberman [8] claimed that virtually every phonetic contrast is cued by several distinct acoustic properties of speech signal and within limits set by the relative perceptual weights and by the ranges of effectiveness of these cues, a change in the setting of one cue can be offset by an opposed change in the setting of another cue so as to maintain phonetic percept. This phenomenon of of perception is generally known as phonetic cue-trading relation. Using the results of new works by Bailey, Summerfield and Dorman [1] , Best, Morrongiello and Robson [2] and Repp [12] , Repp [13] argues that the cue-trading relation operates only in the phonemic mode of perception but not in the auditory mode of perception. On the basis of this point, he claims that speech perception is a special mode of perception different from the mode of perceiving non-speech sounds.

Best et al. [2] investigated the cue -trading relation between silence gap and F1 onset frequency for the "say"/"Stay" continuum as their test stimuli. They found that the cue-trading relation was evident only in the group which was instructed to treat the sinewave analogues as speech sounds but not in the group which was told that the stimuli were non-speech computer sounds. However, their findings provoke further questions about the nature of cue-trading and its relevance to the nature of speech perception. All their test stimuli were confined to the word level. Thus it is necessary to ascertain how the cue-trading relation as a phenomenon behaves at non-linguistic and linguistic levels other than words i.e. sentence, syllable, phonetic and auditory levels, so that the finding may shed more light on the controversy of "speech specificity". The present study was conducted in order to answer the following research questions:

(i) along the five linguistic levels mentioned above, do individuals' (both normal hearing and hearing impaired) cue-trading relations at one linguistic level differ from those of the other levels, with initial voicing contrast as an example?

(ii) if the cue-trading relations differ from one level to another, what is the interlinguistic level pattern of cue-trading?

(iii) how do the answers to the above questions fit into the present controversy of special v. non-special mode of speech perception?

METHOD

Selection of Segments and Creation of Linguistic Levels

The stop consonant type selected for initial voicing contrast was alveolar because alveolar stops have the most confined range of initial F2 and F3 frequencies (see [7] :123). The vowel /a/ was chosen for the syllable level continuum. The diphthong /ai/ (forming the words "dye"/"tie") was chosen for the word level continuum. The same /dai/-/tai/ continuum was chosen for the sentence level as well. In order to create the sentence level processing, the individual steps of the /dai/-/tai/ continuum were placed at the end of PL carrier sentences. The phonetic and auditory level stimuli were the sinewave analogues of the syllable level /da/-/ta/ ten VOT steps.

Syllable Level Stimuli

Using the 12 parameter serial analogue speech synthesiser designed by Clark [3] and [4] , step one of the syllable level continuum, one with 0 VOT i.e. the good /da/ of the /da/-/ta/ continuum, was created first. The frequency values of the three formant patterns of the good /da/ were set after the averaged measurements of the five /da/ spectrograms of five general to broad Australian male native speakers of English. The duration was 300 msec. The fundamental frequency was constant at 125 Hz over the first 85 msec and fell linearly to 90 Hz. The initial formant transitions were stepwise linear and 45 msec in duration. F1 rose from 285 to 770 Hz, F2 fell from 1540 to 1233 Hz and F3 fell from 3019 to 2520 Hz. The duration of the synthesis time

frame was 5 msec i.e. the synthesis data was updated at every 5 msec. Then the VOT continuum for the remaining nine 5-msec steps was created by replacing the periodic voiced (V) excitation with noise and simultaneously increasing the bandwidth of F1 transition to its maximum and hence virtually eliminating the existence of F1 transition . The first /da/-/ta/ continuum created in such a way produced a good/da/ on one end and a good /ta/ on the other. The amplitude levels of the noise and vowel portions (though different in actual measurements) in this continuum were given the nominal 0dB each. Therefore the first /da/-/ta/ continuum can be described as bearing the nominal amplitude pattern of 0dB A (aspiration noise) and 0dB V (vowel portion). Eight more /da/-/ta/ continua were created by increasing and decreasing both the 0dB A and 0dB V amplitudes by 6dB as described below.

A amplitude		V amplitude
+6dB	orthogonally	+6dB
0dB	combined	0dB
-6dB		-6dB

The formulation of these syllable level continua was almost the replica of Repp [11]. Through such an arrangement it was expected that the stimuli with the A=+6dB and V=-6dB pattern would produce more /t/ responses and those with A=-6dB and V=+6dB would produce more /d/ responses. For normal hearing listeners, an increase or decrease in amplitude by 6dB is appropriate to make the stimulus noticeably louder and fainter respectively. Every step from each continuum served as one stimulus. On the test tape all the ten VOT steps of the nine continua (i.e. 90 stimuli) were randomised four times with an interstimulus interval of 3.5 msec and these four sets of randomisation served as four blocks of test stimuli for the syllable level.

Phonetic and Auditory Level Stimuli

The phonetic and auditory level stimuli were the same four blocks of 90 stimuli each from the syllable level. The only difference was that the phonetic and auditory level stimuli were the sinewave analogues of the syllable level stimuli.

Word Level Stimuli

The word level stimuli were in principle the same as the four blocks of 90 stimuli each from the syllable Level. The difference was that the word level stimuli were from the nine /dai/-/tai/ continua instead of the nine/da/-/ta/ continua of the syllable level. With the exception of the initial transition duration, the durations and frequency values of the formant trajectories in the /dai-tai/ continua were set after the corresponding averaged values of the five /dai/ spectrograms of five (general to broad) Australian male native speakers of English. As with the good /da/ stimulus at the syllable level, the duration of the initial formant transitions was 45 msec long in the good /dai/ stimulus. Such an arrangement was necessary in order to maintain the uniformity of VOT steps along the the different linguistic levels. It was the duration of the initial formant transitions which was progressively replaced by noise in order to create VOT steps in this experiment.

Sentence Level Stimuli

The sentence level stimuli were the same four blocks of the word level /dai/-/tai/ ("dye"-"tie" as words) stimuli. In order to create the sentence level processing for the subjects, the "dye"-"tie" test stimuli were presented in the context of the sentences whose semantic, syntactic, vocabulary and phonetic variations were controlled. Every carrier sentence consisted of seven syllables including the stimulus word at the end. On the sentence level test tape, the 90 sentences used were synthesised according to the synthesis by rule system of Australian English by Clark [5] and [6]. The average amplitude of the carrier sentences was maintained at the same value of the nominal 0dB of the V amplitude in the stimulus word. Fourteen spectrograms of the seven basic sentences (each carrying a good "dye" and a good "tie") spoken by Professor Clark were used as norms in synthesising the sentences.

Subjects

Thirteen male and 15 female normal hearing listeners and 7 male and 2female (sensorineurally) impaired listeners took part in the listening tests. They were all native speakers of Australian English and naive listeners of synthesised speech, between 20 and 40 years of age.

Test Procedure

The listening test was conducted in the acoustically treated speech perception laboratory of the Macquarie Speech, Hearing and Language Research Centre (SHLRC). All the listeners wore the Telephonics TDH 49P audiometric headphones with circumaural seals and each listener sat at a test booth. The test tapes were played on a Revox B77 MKII stereo tape recorder. The output level of the tape was controlled by an HH professional power amplifier TPA 25-D with calibration control panel and level meter attached. The tape output level was adjusted in such a way that the amplitude of the loudest sound on the tape was approximately 80dB SPL. The stimuli with the best /d/ and the best /t/ (i.e the A=-6dB/V=+6dB and A=+6/V=-6 amplitude patterns respectively) were used in the anchoring procedure for all the levels. At the sentence level the anchoring was conducted with the original seven basic carrier sentences. For the anchoring at the auditory level the subjects were told to treat one sound (the best/da/ analogue) as sound one and the other (the best /ta/ analogue) as sound two. At the phonetic level, the subjects were told that the sounds were whistled imitations of the /da/ and /ta/ speech sounds. There was a time lapse of at least two weeks between the tests of different levels.

At the four lower levels the task of the subjects was to identify every stimulus (sound one v. sound two, whistled /da/ v. whistled /ta/, /da/ v.

/ta/ speech sounds, and "dye" v. "tie" respectively) and tick their decisions on the sheets provided. At the sentence level the task of the subjects was to write down the whole sentence as they heard it.

METHOD OF ANALYSIS

The /d/ (sound one) responses were counted at every level. The initial data consisted of ten /d/ responses (at the ten VOT steps) for each of the nine continua at every level. Since the role of the +6dB A and V amplitudes as traded cues along the five linguistic levels was more important than the role of the VOT duration, the data were reorganised for every subject in two frameworks i.e. the framework of the role of A and V amplitude levels and the framework of categoricality distance. The following formula was used to reorganise the data for the first framework:

$$\frac{\Sigma\ /d/\ \text{(or sound one) at each continuum}}{\text{number of trials for each stimulus (4)}}$$

If the cue-trading relation was operating as could be expected from the data of previous works (e.g. Repp 1979, Pisoni 1977, Miller et al. 1976 etc.), the continua with the A=-6dB/V=+6dB should attract more /d/ (sound one) responses and those with A=+6dB/V=-6dB should attract less /d/ responses. For the second framework, the framework of categoricality distance, the following formula of data reorganisation was followed.

$$\frac{4 - \Sigma/d/\ \text{of the first five steps} + \Sigma/d/\ \text{of the last five steps}}{\text{total number of VOT steps (10)}}$$

If the responses were strictly categorical, the number of /d/ (sound one) responses for each of the first five steps would be 100% i.e. 4 and that for each of the last five steps would be 0. If the cue-trading is operating, the responses can be expected to be less categorical when the V amplitude decreases and the A amplitude increases. The strength of the first framework of data organisation lies in the comparison of the roles of A and V amplitude levels. The strength of the latter lies in the interlinguistic level comparison of the roles of A and V amplitudes. The reason for this is that the variations in the former were not tied to any fixed value common to all the levels whereas those in the latter were tied to the idealised categoricality of responses.

The analysis of variance with planned contrasts was conducted. The contrasts was planned to ascertain the interlevel comparisons (a. lower two levels v. upper three, b. auditory v. phonetic levels, c. syllable level v. two upper levels and d. word v. sentence levels).

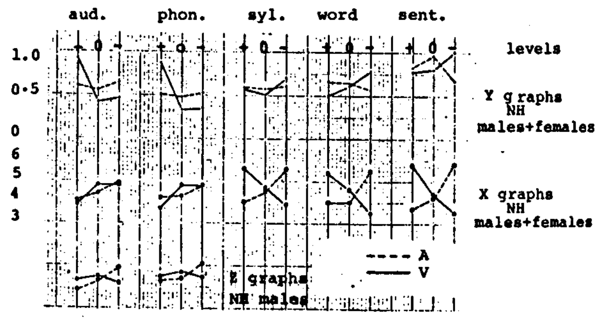


Figure 1 NH Pooled means

SUMMARY RESULTS AND DISCUSSION

In the first analysis the normal hearing (NH) and hearing impaired (HI) group difference was significant at alpha level=.025 (1,33 F.025=5.52 8.57). In the second analysis the NH and HI group difference was significant at alpha level =0.5 (1,33 F.05=4.14 4.88).
In the first analysis, over the grand total of 37 subjects, none of the four contrasts ("a" to "d") was significant. However, interlevel contrast was not the strength of the first analysis. In the second analysis over the grand total of 37 subjects, contrast "b" was not significant but "a", "c" and "d" were (all at alpha level=.025).
This implies that there was no difference in cue-trading relation between auditory and phonetic levels while the cue-trading relations across the upper three levels were different. The pooled means of the A and V amplitude effects for the NH group across the five levels in both the analyses are summarised in figure 1. In the Y graphs of the second analysis, the higher the means move away from the zero line the less categorical are the identifications. In addition, in the Y graphs of the second analysis, the steeper the unbroken lines moves diagonally up from left to right, the more effective is the cue-trading along the V factor and the reverse (left to right downward) is the case with the broken line for the A factor. In the X graphs of the first analysis, the further the means move away from the bar of value 5 (up or down) the stronger is the cue-trading while the steepness of the unbroken and broken lines represents the effect of V and A respectively. In both the analyses, from the syllable level onward, the higher the level, the more linear is the V factor and less linear and more quadratic is the A factor. This suggests that the Va factor exerted a stronger influence than tne A factor at the higher linguistic levels. In the overall analysis it seems that there was no cue-trading relationship at the two lowest levels. However, there was some evidence of cue-trading at the two lowest levels in the HN male group though the extent of such lower level non-linguistic cue-trading is rather insignificant compared with that at the linguistic levels. See Z graphs in figure 1.

CONCLUSION

The results of the experiment indicate that from

the syllable level onward, the higher the linguistic level the stronger the cue-trading and the less categorical were the identifications for both the NH and HI listeners. There was a certain degree of cue-trading at the auditory and phonetic levels though it is not as strong as that at the linguistic levels. The speech mode of perception is special and speech specific in the sense that cue-trading in the speech mode is significantly stronger than that in the non-speech mode, but not in the sense that cue-trading does not exist in the non-speech mode of processing. *

REFERENCES

[1] BAILEY, J.P., SUMMERFIELD,Q.& DORMAN,M. (1977) "On the identification of sinewave analogues of certain speech sounds", Haskins Lab. Status Reports on Speech Research, SR-51/52, 1-25.

[2] BEST, C.T., MORRONGIELLO,B.,& ROBSON,R.(1981) "Perceptual equivalence of acoustic cues in speech and non-speech perception", Percept. & Psycho., 29, 191-221.

[3] CLARK, J.E. (1075) "A 12 parameter serial formant speech synthesiser" , Working Papers of the Speech and Lang. Res. Centre, Macquarie University, January 51-71.

[4] CLARK,J.E., (1976) "Specifications for a 12 parameter formant speech synthesiser", Occasional Papers of the Speech and Lang. Res. Centre, Macquarie University, April

[5] CLARK, J.E. (1979) "Synthesis-by-rule system for Australian English speech", (A doctoral thesis, Macquarie University).

[6] CLARK, J.E,(1981) " A low-level speech-synthesis-by-rule system", Jour. of Phon., 9, 541-497.

[7] FANT, C.G.M., (1973) "Speech sounds and features", (MIT).

[8] FITCH,H.L., HALWES,T., ERICKSON,D.M.& LIBERMAN,A.M. (1980) Perceptual equivalence of two acoustic cues for stop consonant manner", Percept. & Psycho., 27, 343-350.

[9] MILLER, J.D., WIER,C.C., PASTORE,R., KELLY,W. J. & DOOLING,R.J. (1976) "Discrimination and labelling of noise-buzz sequences with varying noise lead times: An example of catergorical perception", Jour. Acoust. Soc. Amer. 60, 410-417.

[10] PISONI ,D.B. (1977) "Identificaiton and discrimination of the relative onset time on two component tones: Implications for the perception of voicing in stops", Jour Acoust. Soc. Amer.,61, 1352-1361.

[11] REPP, B.H. (1979) "Relative amplitude of aspiration noise as voicing cue for syllable initial stop consonants", Lang. & Speech 22,137-189.

[12] REPP, B.H. (1981a) "Auditory and phonetic trading relations between acoustic cues in speech perception: Preliminary results", Haskins Lab. Status Reports on Speech Research, SR-67/68, 165-189.

[13] REPP, B.H. (1981b)"Phonetic trading relations and context effects: New experimental evidence for a speech model pf perception", Haskins Lab. Status Reports on Speech Research, SR-67/68, 1-40

* This research was conducted under the supervision of Professor John Clark, School of English and Linguistics, Macquarie University, Sydney.

The preliminary version of this paper was circulated at the first Australian Conference of Speech Science held at the Australian National University, Canberra, November 1986.

PERCEPTION OF CUES TO A STOP VOICING CONTRAST BY

NORMAL-HEARING CHILDREN AND ADULTS

VALERIE HAZAN LISA HOLDEN-PITT SALLY REVOILE DONNA EDWARD

Dept. of Phonetics and Linguistics
Univ. College London
London NW1 2HE, UK

Sensory Communication Research Lab.
Gallaudet University
Washington, DC 20002, USA

ABSTRACT

The contribution of two acoustic cues, Voice Onset Time (VOT) and vowel onset transitions, to the perception of a /tæd/-/dæd/ contrast was examined for normal-hearing children and adult controls.

INTRODUCTION

The important speech acoustic cues effecting voicing identification of initial stops are voice onset time (VOT) and the vowel onset transitions subsequent to the stop burst [1]. The contribution of these cues for initial consonant voicing perception by adults and children, both normal-hearing and hearing-impaired, has been investigated in various studies [2,3,4,5,6]. However, because these cues have been co-varied in most studies, the relative importance of VOT versus vowel onset transitions for initial stop voicing distinctions remains equivocal. Differences in the stimuli used among these studies may also be a factor in the variations found among these results.

This paper describes an experiment that examined further the use of VOT versus vowel onset transitions for cueing initial stop voicing distinctions by normal-hearing children and adults. Both synthetic and spoken stimuli were tested.

METHOD

Stimuli

Three continua of spoken /dæd/-/tæd/ stimuli and two of synthetic /dæd/-/tæd/ were used as the test syllables. Each continuum comprised eight stimuli among which VOT varied nominally from 18 to 60 ms in 6 ms steps.

Vowel onset transitions were also present in three of the five continua's stimuli. The three continua of natural stimuli were derived from two spoken utterances - a /tæd/ and a /dæd/ -that had been selected for their average acoustic characteristics from a larger pool of syllables [6].

In one continuum, TAD/VOT, a /tæd/ utterance served as a base stimulus; the /t/ burst was appropriately shortened to yield the desired VOT durations for the constituent stimuli of the continuum. The resultant /t/ burst were copied for use in another condition, DAD/VOT. The base stimulus of this continuum was a /dæd/ stem from which the /d/ burst had been removed and replaced by the /t/ bursts of different durations. These same stimuli were used in a third continuum of spoken syllables, DAD/VOT/vowel cutback, but here, the vowel was progressively cut back within most of the continuum to approximate the VOT/transition cue relationship found in natural speech.

The two remaining continua contained stimuli developed via software synthesizer [7]. These were copy syntheses generated to resemble perceptually and acoustically the utterances of /tæd/ and /dæd/ from which the natural continua were developed. The synthetic stimuli contained an initial burst, with major energy peaks at 1620 Hz, 2600 Hz and 4000 Hz. Vowel formant values were not steady state but constituted a best fit to the natural vowel. F1 varied from 500 Hz to 288 Hz, F2 from 1850 Hz to 1535 Hz, F3 from 2650 Hz to 2433 Hz and F4 from 3700 Hz to 3450 Hz. In both continua, the respective VOTs approximated those used in the natural continua. In the "Synthetic TAD/VOT" continuum, F1 started at 500 Hz throughout the stimulus range and contained no initial transition; the F1 onset cue was therefore neutralized. In the "Synthetic DAD/VOT/F1 cutback" continuum, the F1 transition was systematically varied in frequency extent throughout the continuum, with a starting frequency of 400 Hz at the voiced extreme of the stimulus range. The transition duration of F1 was 36 ms.

Subjects

Ten normal-hearing children 7-9 years of age and five normal-hearing adults served as paid listeners. The younger subjects were children of employees at Gallaudet University. All children and adults had pure tone thresholds (3FA mean of .5, 1 and 2 kHz) better than 15 dB HL.

Procedure

The stimuli were presented in single-interval identification trials with "TAD", "DAD" response alternatives. Pictorial sketches and orthographic labels of TAD and DAD were displayed on a touch-sensitive screen used as a response terminal.

The order of stimulus presentation followed a simple adaptive tracking procedure developed for use in perceptual experiments with young children [8]. The average length of test for the children was of 48 trials.

The children were tested during five 30-minute sessions that occurred within a three-week period. The tests were administered to the adults in two sessions of about one hour each. The listeners were tested in IAC audiometric rooms, with the stimuli presented monaurally through a TDH-39 earphone (MX 41/AR cushion) in a headset. Stimuli were presented at 75 dB SPL. Stimulus presentation and response tallies were under computer control (DEC PDP-11/23 and 11/34).

RESULTS AND DISCUSSION

For each listener, mean results were computed over the four repetitions of each continuum (three repetitions for the DAD/VOT continuum). A Maximum Likelihood technique [9] was used to fit a cumulative normal function to each set of data. Two measures of performance were derived: the phoneme boundary (50% labeling point of the fitted curve) and the gradient of the identification function (slope). The results are summarised in Table I for the two groups of listeners individually and combined. For each performance measure, analyses of variance were carried out using factors: groups (adults versus children) by conditions (the five stimulus continua) treated as repeated measures.

Phoneme Boundaries

The listener groups did not differ for the phoneme boundaries measured. The group of adults and of children obtained similar /d/-/t/ phoneme boundaries [$F(1,13)=.02$, $p=0.9$], and showed no interaction with the test conditions [$F(4,52)=.9$, $p=0.5$]. This outcome would suggest that the general age difference between the two listener groups was insignificant with respect to their use of the VOT and vowel onset cues for locating /d/-/t/ phoneme boundaries.
The statistical similarity between the groups for phoneme boundary enables their results to be combined for testing differences among conditions. An additional analysis of variance carried out for the total group of listeners revealed that the effect of test condition was significant [$F(4,56)=7.8$, $p=0.0$]. To determine which conditions contributed to this effect, Tukey's test of honestly significant differences (hsd) was applied. For the natural stimuli, the phoneme boundary of the TAD/VOT continuum was significantly shorter than that for other conditions of natural and synthetic stimuli [hsd, $p<.05$]. The phoneme boundary obtained for the synthetic version of TAD/VOT was also shorter than that found for the synthetic continuum with F1 onset cues, DAD/VOT/F1 cutback. However, this result fell just short of statistical significance. These findings indicate that the continua lacking spectral cues to the voicing contrast in the vowel onset required a shorter VOT to be perceived as /dæd/ than stimuli which contained these cues. Others have reported similar results with normal-hearing [2] and some hearing-impaired children [5]. Overall, however, VOT appears to prevail over cues in the vowel onset for effect on initial stop voicing perception. Indeed, categorisation of the stimuli was achieved despite conflicting spectral cues in the TAD/VOT and DAD/VOT conditions. In a recent study, Revoile et al. [6] found that the insertion of aspiration between voiced stop transients and subsequent vowels in spoken stimuli yielded a near complete reversal in perception from voiced to voiceless.

Note that between the natural and synthetic stimuli, conditions with analogous cues (i.e. TAD/VOT, and also DAD/VOT/cutback) yielded similar results. However, Table I reveals that the synthetic stimuli effected greater standard deviations for phoneme boundary means than those found for the natural stimuli. Also, among the continua for natural stimuli, one of the condition with conflicting voicing cues (F1 transition in the presence of long VOTs), DAD/VOT, produced larger standard deviations than found for the other two conditions. We may speculate that the greater variability in results for the synthetic condition and natural with conflicting cues was due to a more artificial quality inherent to these stimuli. Statistically, however, the spoken stimuli and synthetic stimuli generated to resemble the spoken stimuli produced similar perceptual effects for distinction of the voicing contrast in initial alveolar stops, at least for these normal-hearing listeners.

Identification function gradients

A significant difference in identification function gradient was obtained between the two listener groups [$F(1,13)=15.32$, $p=.002$]. Table II reveals that the gradients for the adult group are steeper than those for the children in each condition. This outcome suggests that the children were more tentative in their phoneme distinctions of /d/ vs /t/ than were the adults. This effect is also reported by Simon and Fourcin [2] who found that age-related development in the ability to label voicing contrasts was mirrored by an increase in identification function gradient. An interaction was found between listener group and test condition [$F(4,52)=3.22$, $p=0.2$]. Examination of the means shows that this tendency is largely due to the considerably steeper slope observed for the adults in the natural TAD/VOT condition. This effect is not found for the children's group where very little difference is observed with respect to the function gradient among the five conditions.

When the two listener groups are examined separately for condition effects, neither show a significant difference in function gradient among conditions. However, for the adult group, a greater distinction in function slope is observed among conditions [$F(4,16)=2.37, p=.1$] than is seen for the children's group [$F(4,36)=.69, p=.6$]. Large standard error measures were obtained for identification function gradients for the natural and synthetic TAD/VOT and natural DAD/VOT showing greater inter-individual variability in conditions with conflicting spectral cues.

CONCLUSION

Results confirm the primary importance of the temporal VOT cue over the spectral vowel onset cue to the voicing contrast in initial plosives. Vowel onset characteristics were however shown to have a clear secondary effect, as shown by a shift in boundary, when the cue is absent, in both children and adults. Although children gave very similar labeling to edited natural stimuli than adults, they seemed less affected by a removal of vowel onset cues.

High quality synthetic speech did provide a good match to results obtained with natural edited stimuli, for both adults and children. However, greater inter-individual variations in labeling were found both for adults and children. As a result, the shift in boundary between the TAD/VOT and DAD/VOT / cutback conditions, which had been strongly significant in the natural edited stimuli was found to be short of statistical significance using synthetic stimuli. This would suggest that edited natural stimuli, by providing more homogeneous results, may be more reliable than synthetic stimuli in cue weighting experiments. However, there are limitations in the types of cues which may be altered through computer processing of natural speech, so that synthetic speech does still provide the greatest flexibility when constructing stimulus continua in which spectral rather than temporal patterns are varied.

REFERENCES

[1] Lisker,L. & Abramson,A.S. A cross-language study of voicing in initial stops: acoustic measurements. Word, 20, 384-422, 1964

[2] Simon,C. & Fourcin,A.J. Cross-language study of speech pattern learning. J.Acoust.Soc.Am., 63, 925-935, 1978

[3] Parady,S., Dorman,M., Whaley,P. & Raphael,E. Identification and discrimination of a synthesised voicing contrast by normal and sensorineural hearing-impaired children. J.Acoust.Soc.Am., 69, 783-789, 1981

[4] Johnson,D., Whaley,P. & Dorman,M.F. Processing of cues for stop consonant voicing by young hearing-impaired listeners. J.Speech Hear.Res., 27, 112-118, 1984

[5] Hazan,V. & Fourcin, A. Microprocessor - controlled speech pattern audiometry: preliminary results. Audiology, 24, 325-335, 1985

[6] Revoile S., Pickett, J.M., Holden-Pitt,L.D., Talkin,D., Brandt,F.D. Burst and transition cues to voicing perception for spoken initial stops by impaired and normal-hearing listeners. J.Speech Hear.Res, 30, 3-12, 1987

[7] Klatt,D.H. Software for a cascade/parallel formant synthesizer. J.Acoust.Soc.Am., 67, 971-995, 1980

[8] Hazan,V. Speech pattern audiometric assessment of hearing-impaired children. Doctoral Dissertation. University College London, 1986

[9] Bock,K.D. & Jones,L.V. "The Measurement and Prediction of Judgment and Choice". Holden-Day, San Francisco, 1968

	TEST CONDITIONS				
	Natural stimuli			Synthetic Stimuli	
Phoneme Boundary	TAD/VOT	DAD/VOT/ Vowel Cutback	DAD/VOT	TAD/VOT	DAD/VOT/ F1 Cutback
Children n=10	28.7 (2.2)	32.4 (1.3)	32.3 (3.2)	29.6 (5.2)	31.9 (3.4)
Adults n=5	26.1 (1.7)	31.5 (3.1)	32.7 (3.1)	30.1 (4.5)	33.3 (5.6)
Total n=15	27.8 (2.3)	32.1 (2.0)	32.4 (3.1)	29.8 (4.9)	32.3 (4.1)

Table I : Mean phoneme boundary values (in milliseconds)

ID Function Gradient (Slope)	TAD/VOT	DAD/VOT/ Vowel Cutback	DAD/VOT	TAD/VOT	DAD/VOT/ F1 Cutback
Children n=10	-.92 (.52)	-.82 (.23)	-1.03 (.50)	-.88 (.36)	-1.04 (.60)
Adults n=5	-2.34 (.74)	-1.52 (.33)	-1.53 (.68)	-1.28 (.53)	-1.31 (.74)
Total N=15	-1.40 (.90)	-1.05 (.43)	-1.20 (.59)	-1.01 (.45)	-1.13 (.63)

Table II : Mean identification function gradients

THE ROLE OF INTENSITY IN BREATHY VOICED STOPS: A CLOSE LINK BETWEEN PRODUCTION AND PERCEPTION

LIESELOTTE SCHIEFER

Institut für Phonetik und Sprachliche Kommunikation
der Ludwig-Maximilians Universität München, FRG

ABSTRACT

The acoustic analyis of the intensity difference between the breathy and steady vowel portions following a breathy voiced stop in Hindi revealed a significantly lower intensity in the breathy portion compared with the steady one with a clear influence of the vowel. The perceptual importance of that acoustic cue was tested with 4 CV combinations. The results reflect an interaction between the intensity and vowel on the one hand, and an interaction between intensity, amplitude of H1/H2, and F1 on the other hand.

INTRODUCTION

Over the last few years several studies on the difference between breathy voiced or murmur phonation and normal voicing were carried out with languages such as Gujarati [1], !Xõo [1, 3], Hmong [2], Hindi [4]. All studies confirmed that the amplitude of the first harmonic (H1) is always higher in murmur (henceforth breathy) phonation compared to the second harmonic, resulting from the sinusoidal glottal source waveform. The perceptual importance of this acoustic parameter was shown by Bickley [1]. On the other hand, breathy voiced stops in Hindi are always accompanied by a drop in overall intensity after the release of the stop. We have already examined the perceptual importance of this acoustic cue using the method of speech editing for stimulus generation and a naturally produced syllable /ḍho/ as point of departure. The results were in agreement with the assumptions of categorical perception [4]. As these results were obtained with only one CV combination, it seemed interesting to test the perceptual load of the acoustic cue with other vowels, too, and to compare these results with acoustic data.

MATERIAL AND INFORMANTS

For the acoustic analysis a list of words was prepared, which contained the breathy stops /bh dh ḍh gh/ in word-initial position. Each stop was followed by the phonemically long vowels /a e i o u/ and occurred in 10 different words, which consisted of either one, two, or three syllables. The material was not controlled for the consonant following the initial CV syllable. Twenty lists were prepared, each containing a subset of the words in randomized order. The lists were read by three informants (1 female, 2 males, aged 23 to 40), all native speakers of Hindi, originating from New Delhi or Uttar Pradesh. The recordings were made in New Delhi in the language lab of the Centre of German Studies, School of Languages of the Jawaharlal Nehru University using a Uher Report and a Senheiser MD421N microphone. The distance to the microphone was set at about 50 cm. The same word list was recorded from another informant (35 years, female) in Munich in the soundproofed room of the Institute using a Telefunken M15 tape recorder and a Neumann U87 microphone. This recording served as font for the manipulation and generation of the stimuli employed in the perception tests.

PROCEDURE

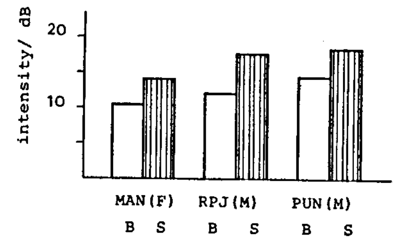
The material was digitized on a PDP11/50 with a sample rate of 20 kHz, filtered with a cut off frequency of 8 kHz, and stored for further analysis. The material was manually segmented with the help of a segmentation routine (for further information cf. [5]). Four different parts in the initial CV syllables were defined: voicing lead, burst + voiceless aspiration, breathy part of the vowel, and steady part of the vowel. All periodic portions were segmented into single pitch periods [5]. For the acoustic analysis the intensity was calculated for each pitch period and averaged over all periods of the breathy and steady vowel portion, respectively, for each speaker. Separate analyses of variance were applied to all comparisons of means for all speakers. The normality of distribution was checked by a chi-square procedure, while homogeneity of variance was controlled for by applying the chi-square statistics for independent measurements. The level of significance was set to p < .05. Multiple comparisons of means were calculated by the use of an a priori F statistic.

For the perception tests 4 CV combinations /dha dho dhu dhi / were selected for manipulation in order to test the interaction between intensity and the vowel. A set of programs was used to generate the stimuli. The procedure has been described in detail elsewhere [4, 5]. The first CV syllable was separated from the rest of the words and the breathy portion of the vowel was eliminated totally. The resulting syllable consisted of voicing lead and burst (which remained unchanged), and the steady portion of the vowel. The fundamental frequency of the vowel was adjusted to 210 Hz for all CV combinations, with a rise over the first five pitch periods and a fall over the last five periods. The first stimulus of each continuum was generated by superimposing a quasi-linear intensity curve on the first 21 pitch periods, the first period being adjusted to 25 dB, the 21st to 55 dB. The intensity was kept constant for the rest of the vowel, with a decrease over 5 periods at the end of the contour. The other 6 stimuli of each continuum were derived from the first stimulus by increasing the intensity onset in the 1st period by 5 dB. For each CV condition identification and discrimination tapes were prepared. In the identification test each stimulus occurred 5 times in randomized order with a pause of 3.5 sec after each stimulus and a pause of 10 sec after a block of 10 stimuli. For discrimination the AX paradigm was used with the step size = 2. Both presentation orders AB and BA as well as AA occurred. The interstimulus interval was 500 ms, pairs were separated by 3.5 sec, blocks of 10 pairs by 10 secs. Each pair occurred 3 times in randomized order. Answer sheets were prepared to allow responses for either breathy or voiced stops in a forced-choice paradigm in the identification task, whereas in the discrimination task subjects had to decide whether the stimuli within a pair sound the "same" or "different". All perception tests were run in the language lab of the Centre of German Studies in New Delhi using a Telefunken language trainer with head phones. The tests were run at a comfortable listening level. About 15 subjects participated in the tests. All were staff or students of the School of languages and were paid for their participation.

RESULTS

Acoustic analysis. Fig. 1 displays the acoustic results for the three speakers averaged over all stops and vowels. Fig. 2 to 4 display the results for the vowel, Figs. 5 to 7 for the stop conditions separately for the three speakers. It is obvious that the intensity of the breathy vowel portion differs significantly from that of the steady portion in all speakers. On the other hand, the amount of difference between both. portions is not the same in all subjects: it is large for RPJ (11.1 dB), and smaller for PUN (7.9 dB) and MAN (7.22 dB). The influence of the vowel on the intensity is large for all informants: MAN F(4,177 = 18.821; p < .001; RPJ : F(4,185) = 7.65; p < .001; PUN: F(4,184) = 14.711; p < .001). But the vowels do not contribute in the same way to the intensity difference between the breathy and steady portion as can be seen from the following diagrams, which show the significance between the single vowels.

Fig. 1: Intensity of the breathy and steady vowel portions in dB averaged over all stops and vowels; plotted separately for the speakers MAN(F), RPJ(M), and PUN(M)



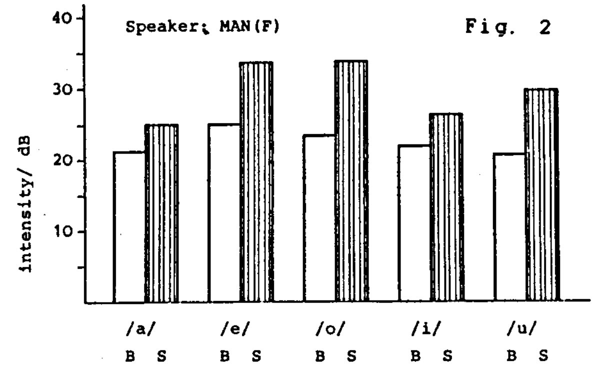

```
MAN(F)   a   i   e   u   o
         x...x   x   x...x
RPJ(M)   u   a   i   o   e
         x.......x   x...x
PUN(M)   i   u   e   a   o
         x...x       x   x
             x...x
```

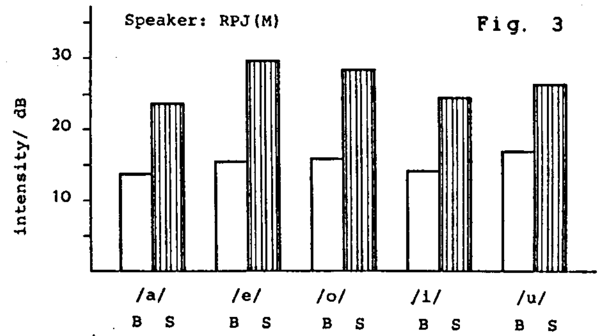
(The diagrams should be read as: vowels underlined by a common dotted line do not differ significantly, whereas vowels not underlined by a common line do.)
MAN shows more influence of the tongue position on the intensity difference as it is largest for the back vowels /o u/, whereas the influence of the tongue height plays an important role in RPJs productions: mid vowels have the larger differences. The results from PUN are not clear, as /o/ and /a/ produce the largest, high vowels the smallest intensity difference. If summarized over all informants the following rank order appears:

i	u	a	e	o
6	7	7	11	14

In other words: the intensity difference between the breathy and steady portion (of the vowel) is a function of the tongue height of the vowel: /i u/ < /a/ < /e o/. The influence of the stop's place of articulation is less compared with the vowel, as only MAN shows a significant influence: MAN: F(3,195) = 10.322; p < .001 ; RPJ: F(3,208) = 2,1; p > .05; PUN: F(3,184) = .854; p > .05.

Perception tests. The results from the identification task are plotted in Fig. 8. The number of participants is given in the figure. It is obvious that subjects did divide the continuum into two parts only in the /dho/ condition, which resembles best





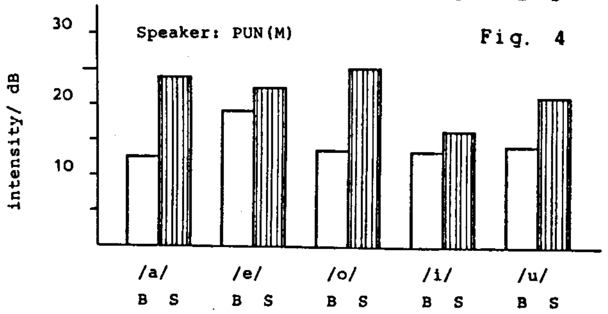


Fig. 2 to 4: Intensity of the breathy and steady vowel portions as a function of the vowel plotted separately for the three speakers

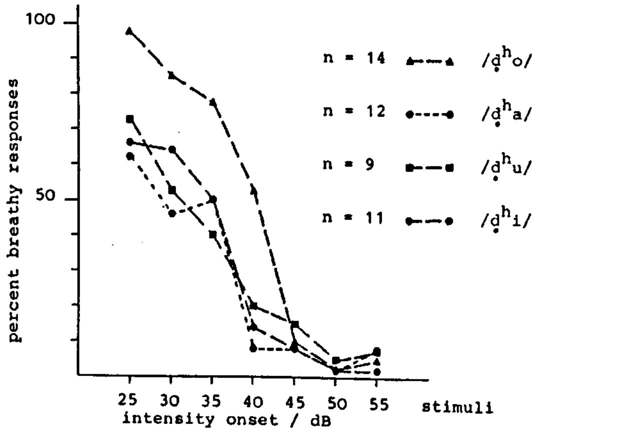
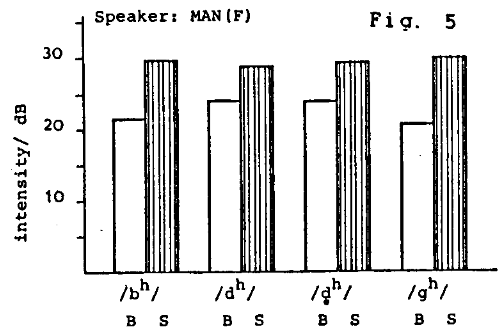
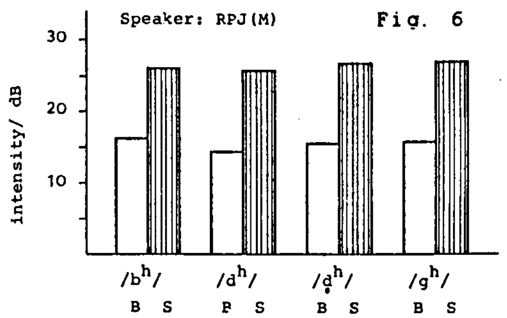


FIG. 8: Percentage breathy responses displayed for the four vowel conditions /dha dho ḍhi dhu/





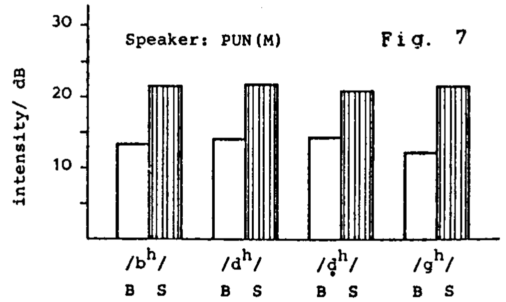


Fig. 5 to 7: Intensity of the breathy and steady vowel portion as a function of the stop plotted separately for the three speakers

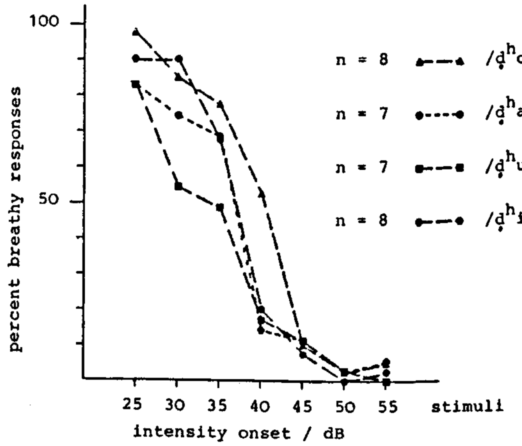


Fig. 9: Percentage breathy responses displayed for the four vowel conditions /dha dho dhi dhu/

the assumptions of categorical perception. In the other tests only the first stimulus was assigned to the breathy category. We asked if these results may be due to an interaction between intensity and the vowel, or if they reflect a difference in subjects ability to make use of that special acoustic cue. Therefore we reanalyzed the results and included only those subjects in the analysis who unambigously assigned two categories to the continuum. These results are given in Fig. 9. This time, the identification function improoved for all conditions. The boundary between breathy and voiced occurs latest in the /ḍho/ condition (cf. Table 1), earlier in /ḍha/ and /ḍhi/, earliest in /ḍhu/ where the boundary is less steep. The main effect of continua is significant ($F(3,26) = 3.644$; $p < .05$), but on the other hand, the continua do not differ significantly from each other as shown in the diagram:

```
ḍho   ḍhi    ḍha    ḍhu
x.......x    x.......x
      x...x
```

TABLE 1: Points of intersection between the identification function and the 50% line.

/ḍho/	/ḍhi/	/ḍha/	/ḍhu/
4.06	3.34	3.26	2.68

The results from the discrimination tasks correspond well with those predicted from the identification task using the Haskins formula. In all tests the discrimination peak of the obtained discrimination function correspond in location and height with the calculated one. On the other hand, subjects could discriminate slightly better than predicted, but the difference was not significant.

DISCUSSION

The results from the acoustic analysis confirm that vowels contribute in different degrees to the intensity difference between the breathy and steady vowel portions after breathy voiced stops. This difference is largest for the mid vowel /o/ in all speakers examined. On the other hand, the perception tests gave best results in the /ḍho/ condition which showed the steepest boundary and the latest intersection between the identification function and the 50% line. This means, that the back vowel /o/ needs less intensity difference in perception than do the high and low vowels. At first glance, one would like to explain these results by a close link between the production and perception: the acoustic cue "intensity" is most powerful when it is applied to mid vowels, less powerful with other vowels. In order to make sure that no other acoustic cue was involved, we reexamined the acoustic structure of the syllables used for manipulation with regard to the relationship between the amplitude of the first and second harmonic in the so called steady portion of the vowel, where no breathiness could be detected auditorily. The results are as follows:

/ḍho/ : H1 < H2
/ḍhi/ : H1 > H2
/ḍha/ : H1 < H2
/ḍhu/ : H1 > H2

This means, that the amplitude of H1 is higher in the high vowels /i u/, whereas H2 exceeds the value of H1 in either /o/ or /a/. These relationsships are undoubtedly due to the formant structure of the vowels, where F1 interacts with H1 in the high vowels, and with H2 in /o/. No interaction between F1 and H1/H2 occurs in /a/. We believe that our results reflect an interaction between the overall intensity, the amplitude of H1 and H2 as well as F1. As these results were rather unexpected, further investigations are needed to explain the extent of that interaction.

REFERENCES

[1] Bickley, C.: Acoustic analysis and perception of breathy vowels. Working Papers Vol. 1, MIT Speech Communication, pp. 73-83 (1980)

[2] Huffman, M.K.: Measures of phonation types in Hmong. University of California Working Papers in Phonetics 51:1-25 (1985)

[3] Ladefoged, P.: The linguistic use of different phonation types. University of California Working Papers in Phonetics 54: 28-39 (1982)

[4] Schiefer, L. - Kotten, K.: Amplitude envelope and the perception of breathy stops in Hindi. Proc. 10th Int. Congr. Phon. Sci., Utrecht, pp. 459-463, (Cinnaminson, Dordrecht, 1984)

[5] Schiefer, L.: Fo in the production and perception of breathy stops: evidence from Hindi. Phonetica 43: 43-69 (1986)

Acknowledgment

This research was supported by German Research Council (DFG) grant Schi 181/2-1. I am grateful to Prof. R.P. Jain and Prof. P. Talgeri for their help in selecting the participants for the perception tests and the permission to use the Language Lab, as well as Swaran Thakur-Weifenbach, Manorama, R.P. Jain, and Puneet who served as informants.

INHERENT VOWEL DURATION IN RUSSIAN: PRODUCTION AND PERCEPTION DATA

VLADIMIR B. KUZNETSOV
Moscow State Institute of Foreign Languages
Moscow, USSR 119o34

ARVO OTT
Dept. of Computer Control
Institute of Cybernetics
Tallinn,Estonia,USSR 2oo1o8

ANATOLY V. VENTSOV
Pavlov Institute of Physiology
Leningrad,USSR 199164

ABSTRACT

In this paper the quantitative data concerning inherent vowel duration in Russian are presented. The established duration differences are estimated from the point of view of their perceptual significance in an experiment on lexical stress location.

INTRODUCTION

It is a well established fact that all else being equal, duration of stressed vowels is determined by the characteristics of the corresponding vowel gesture: the higher the tongue, the shorter the vowel. To be more specific, vowel duration is considered to be a result of superposition of jaw movement on the opening and closing gesture of the lips [1]. Thus, the vowel of a given phonetic identity has its own characteristic duration - inherent vowel duration (IVD).

Although in Russian IVD phenomenon has been studied by many researchers [2,3,4] the question whether the observed duration differences have any significance either at the production or perception level is still opened. This is mainly due to the fact that most results of measurements are qualitative. Among the other reasons, poor control of the experimental conditions, interpretation of the data in terms of different phonetic categories (phonemes vs. allophones), unjustified averaging over the speakers and phonetic contexts have to be mentioned.

The aim of the experiments reported in this paper was to gather statistically reliable and valid data on IVD and to assess the perceptual importance of the determined IVD differences.

EXPERIMENT 1:INHERENT VOWEL DURATION (IVD)

Vowel durations were measured in a monosyllable of CVC type, spoken as word in an identical sentence frame "Say....again". Palatalized or nonpalatalized fricative [s] was used to form a symmetrical environment. Each of four speakers (two male ones: M1,M2 and two female ones: W1,W2) recorded a list of 330 sentences (10 vowels * 33 repetitions). To achive constant speech rate thoughout the recording session the speaker was asked to synchronize the onset of the sentences with a periodic light pulse. The phonetic identity of the test vowels was checked up by 8 listeners during an identification experiment.

Three segmentation procedures were used to measure IVD. The beginning and end of a vowel were recognized: (1) by the onset/offset of voicing; (2) by the offset/onset of high-frequency noise; (3) by sharp minimums on the amplitude curves.

There was a good agreement among the three sets of measurements. Taking into consideration the fact that the third segmentation procedure produced the least scatter of measurements, only the data obtained by this procedure were subjected to futher analysis.

The amplitude envelope was obtained by processing the tape recordings of the speech material through a Brüel&Kjaer graphic level recorder operated at a 100 mm/sec paper - and 1000 mm/sec writing-speeds with a high frequency preemphasis. To increase time resolution the play-back speed of the tape-recordings was reduced twice. Preliminary spectrographic analysis of the test words revealed that the minima on the amplitude envelope coincided with the onset of the voicing of the vowel, on the one hand, and the rapid energy decrease in the frequency region of the second and higher vowel formants and the onset of the frication noise, on the other. It is commonly agreed, that these acoustic cues provide reproducible and valid boundaries for the measurement of vowel duration [5].

Mean vowel duration is regarded as an estimate of IVD. The four data matrics (10*33) were submitted to the following statistical analysis: (1) to determine the effect of tempo changes a one-way analysis of variance was carried out (to test the null hypothesis that the mean vowel duration of the 33 consecutive groups of ten different vowels are the same); (2) similar technique was used to assess significance of IVD differences; (3) T-method of multiple comparisons [6] was applied to determine the critical value of IVD difference; (4) to test the significance of the proposed classification of vowels according to their IVD the duration data was subjected to S-method of multiple comparisons [6]. All statistical tests were conducted with alpha=0.05. More detailed description of the experimental method and statistical procedures is presented in [7].

Results of the analysis of variance have shown that the speaking rates were kept constant and the differences in vowel durations were statistically significant. Critical IVD difference was within 6-7 msec. By means of the method of multiple comparisons the vowels were reliably rank-ordered and subdivided according to their IVD values into the following classes: {ӓ,a}, {ӧ,ӭ,э,о,ӱ} and {у,ы,и} (cyrillic characters are used to symbolize the vowels, the dots above letters indicate the allophones in the context of palatalized consonants). Typical duration ratio of the IVD means in the classes is 1.00:0.90:0.75. The IVD patterns of the four speakers are presented in Fig.1.

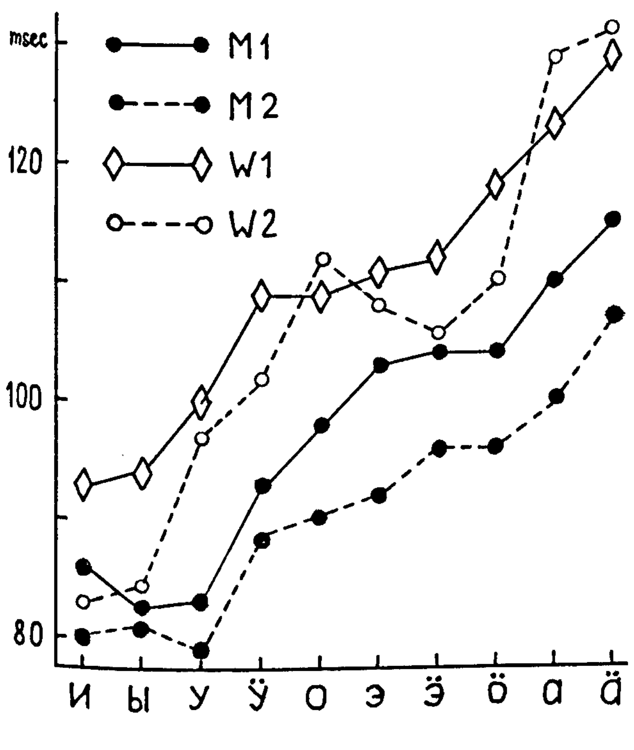


Fig.1 IVD pattern of the four speakers

EXPERIMENT 2: RANGE OF IVD VARIABILITY

To estimate the range of IVD variability two types of speech material were used: (1) the same CVC syllables as in experiment 1, but spoken in isolation, (2) five-syllable nonsense word [bist<V'>tfurnaja] embedded in the carrier phrase "I was told that the oldest ... company had gone bankrupt". The immediate consonantal context of the test vowel was palatalized or non-palatalized. Speaker M1 was instructed to read the materials at his normal speech rate and not to insert any pauses into the carrier sentence. The arrangement of the speech material, recording conditions and segmentation procedure were similar to those described above. A more detailed description of the experiment may be found in [8]. Mean vowel durations measured in the two contexts are regarded as maximum and minimum IVD estimates respectively. Maximum and minimum IVD along with the normal ones averaged across vowels belonging to the same class are presented for speaker M1 in Table 1. Examination of the table reveals that our data on minimum IVD support the suggestion put forward in [9] that vowel incompressibility is relative to its inherent duration.

Table 1. Mean IVD for vowel classes (in msec)

IVD	classes ӓ,a	ӧ,ӭ,э,ӱ	у,ы,и	IVD ratio
normal	113	101	84	1.00:0.90:0.75
maximum	167	152	142	1.00:0.91:0.85
minimum	90	75	63	1.00:0.84:0.71

Experimental data reported in this paper provide some evidence against the concept that in Russian the degree of opening is a single factor determining vowel duration. One can hardly explain, for examle, within the framework of Lindblom's model of lip-jaw coordination [1] why the vowels [и] and [э] that are produced by quite similar articulatory gestures, judging from the corresponding F-patterns [13], differ so much in IVD, and the vowels [ӱ,ӧ,ӓ] are systematically longer than [у,о,а], though it is recognized that [ӓ] is closer than [а]. Futher research is needed to clarify the significance of these findings.

Nevertheless, it may be concluded that the acoustically-defined IVD is characteristic of Russian as well. But still there is a possibility that the auditory system of human being uses quite different segmentation criteria for the measurement of subjective duration and the differences in IVD might be already neutralized at the stage of measurements. Evidently, the answer to this question may be obtained only by investigation of speech perception.

EXPERIMENT 3: PERCEPTUAL ROLE OF IVD

The data on the perceptual role of IVD are rather contradictory: on the one hand, there is an evidence that "naturalness" of synthetic vowels is increased if an appropriate IVD pattern is used to control the duration parameter [10], but, on the other hand, the results reported in [11] indicate that IVD is not important for the perception of stress.
Since in Russian the vowel duration is known to signal the position of word stress [12], the mechanism of stress perception has to take into account IVD which must have effect on the results of psychoacoustic experiments with the vowel of different identity.
In the experiment described below natural russian words were used as stimuli. Most of the words were disyllabic with an open final syllable. The first vowel of the words was either [и] or [у], the second one was always [a]. It should be noted that the spectral properties of these vowels do not change appreciably in the pre- and post-stressed positions. The words formed phonetic minimally contrastive pairs, differing only in the position of stress, for example, "ти'хо-тихо'", "у'хо-уха'".
The natural vowels were replaced in the words by semi-synthetic ones of required duration. The method of stimulus generation and experimental procedure are described in full detail in [14]. Duration of the first vowel was 11 fundamental periods (one period was 8.7 msec). Duration of the second vowel varied from 7 to 23 periods.
In Fig.2 and 3 the frequency of response "the second vowel is stressed" is plotted as a function of its duration for the vowel pairs [и-а] and [у-а] respectively. The lines designated by opened circles represent the data when the amplitudes of the vowels in the pair were made equal, the lines marked with crosses represent the data when the effective values of the vowels were equalized. For the sake of comparison the results of the experiment with words comprising identical vowels are also presented (light lines without special signs).
Let us assume that the vowels having the same subjective duration, are judged as stressed with equal propability, then from the results displayed in Fig.2 and 3 it follows that the vowel [a] must have a longer acoustic duration to be perceived as subjectively equal to the vowels [и] and [у]. Since the test vowels were produced by repetition of one fundamental

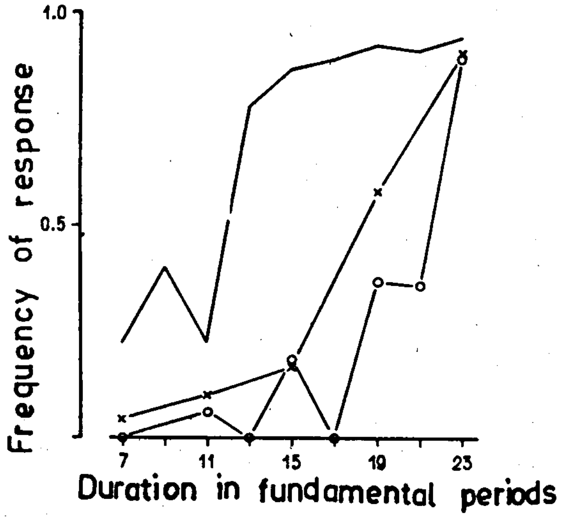


Fig.2.

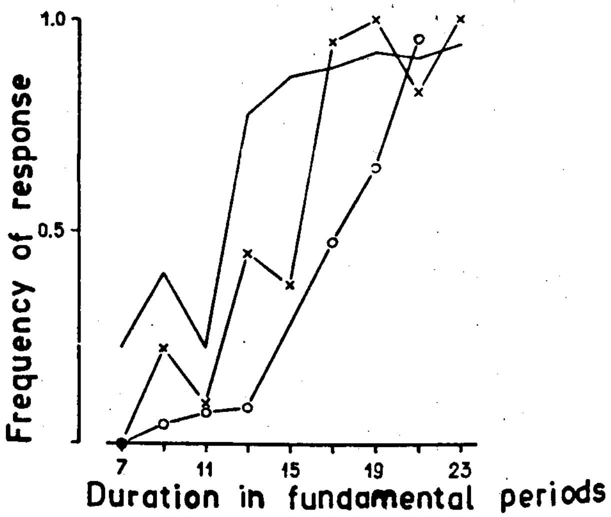


Fig.3.

period singled out from the steady-state part of the corresponding natural vowel [14], the established discrepancy between the acoustic and subjective durations can not be ascribed to the segmentation. Consequently, this discrepancy may be considered to be a result of the IVD effect on the perception of stress when the judgement of stress is based on vowel duration. Thus, the reality of IVD in Russian has been demonstrated not only at the production but at the perception level as well.

This raises an interesting problem of establishing formal rules concerning IVD that may be used in the algorythmic descriptions of word stress perception.

REFERENCES

[1] B.Lindblom "Vowel duration and a model of lip-mandible coordination", STL-QPSR,4,1967
[2] L.V.Shcherba "Qualitative and quantitative description of russian vowels" (in russian), Peterburg,1912
[3] L.R.Zinder "Effect of speech rate on production of some sounds" (in russian) In:"Phonetic problems", Leningrad, 69, p.3-27, 1964
[4] K.Bolla "Some problems of vowel duration ratio in Russian" (in russian), Bulletin of Moscow State University, Philology, Moscow, 3, p.54-62,1968
[5] E.Fisher-Jorgensen, B.Hutters "Aspirated stop consonants before low vowels, a problem of delimination" Ann.Rep.Inst.Phon.Univ.CPH, Copenhagen, 15, p.77-102, 1981
[6] G.V.Glass, J.C.Stanly "Statistical methods in education and psychology", New Jersey, 1970
[7] V.B.Kuznetsov, A.Ott "Inherent vowel duration in Russian" Estonian Papers in Phonetics, Tallinn, p.67-95, 1987
[8] V.B.Kuznetsov, A.Ott "Speech synthesis by rule. Algorithm of phonetic transcription and control of segmental duration" (in russian), Tallinn, 1987 (in print)
[9] D.H.Klatt "Interaction between two factors that influence vowel duration" Journ.Acoust.Soc.Am., 54, p.1102-1104, 1973
[10] R.Petersen "The influence of tongue height on the perception of vowel duration in Danish" Ann. Rep.Inst. Phon.Univ.CPH., Copenhagen, 8, p.1-10, 1974
[11] E.Rosenvald "The role of intrinsic FO and duration in the perception of stress" Ann.Rep.Inst.Phon.Univ.CPH., Copenhagen, 15, p.147-166, 1981
[12] N.I.Jinkin "Perception of stress in Russian words" (in russian) Trans. of the Academy of pedagogical sciences, Moscow, 54, p.7-83, 1954
[13] V.B.Kuznetsov, A.Ott "Spectral properties of russian stressed vowels in the context of palatalized and nonpalatalized consonants" (contribution to the present congress), Tallinn, 1987
[14] A.V.Ventsov "What is the reference that sound durations are compared with in speech perception?", Phonetica, 40, p.135-144, 1983

ON THE MULTIPLICITY OF FACTORS AFFECTING P-CENTER LOCATION

BERND POMPINO-MARSCHALL HANS G. TILLMANN

Institut für Phonetik und
Sprachliche Kommunikation
Universität München, FRG

ABSTRACT

In a series of experiments using the technique of metronome speech it could be shown that the p-center location in production is not solely dependent on the duration of the segments of the uttered syllables but on context and on position within alternating syllable sequences as well.

INTRODUCTION

Generally, it is assumed that the location of the p-center - the psychological moment of syllable onset - of monosyllables is solely dependent on the duration of the initial consonant(s) and that of the syllable rhyme. This should hold true for perception [2] as well as for production [1] As a description of the results of his perception experiments Marcus [2] gives the following formula:

$$P = .65 * C + .25 * VC + const.$$

where P is the p-center location measured relative to the acoustical beginning of the syllable, C the duration of the syllable-initial consonant, VC the duration of the syllable rhyme, and const. a constant. The following experiments were run to test the predictions of this formula in production of systematically varied monosyllabic material.

GENERAL METHOD

In a series of production experiments in which subjects had to utter sequences of the same or alternating monosyllables in beat with a computer generated metronome signal we measured the position of the beginning of the metronome signal relative to the onset of the the vowel as indicator of relative p-center location.

FIRST SERIES OF EXPERIMENTS

The material of the first set of experiments was of the form /C+ak/ (C = /p/, /b/, /f/, /v/ or /m/; in experiments 3, 5-8 also /ʃp/ or /ʃm/). The monosyllables were uttered in sequences composed of two alternating syllables or as homogeneous sequences of seven repetitions of one single syllable in beat with the metronome presented via headphones. At the same time the utterances of the subject were digitally recorded at a sampling rate of 20 kHz and the beginning of the DA-output of the metronome signal was marked in the digitized input for later processing of the data. Measurements of metronome beginning relative to the acoustic signal and of segment durations were made from the oscillogram trace using the speech editing program at the institutes PDP 11/50. The measurements were made for the second till sixth syllable of all sequences. The experiments were run with two experienced male subjects (odd numbered experiments: subject one; even numbered: subject two). For the first two experiments we used a metronome rate of 60 beats per minute with a 5 msec 3-kHz tone burst as metronome signal, the second set was run with a rate of 90 and the same metronome signal. The third set of experiments was also run at a rate of 90 but with the syllable /vak/ as metronome signal.
The metronome position data were analysed by two-factorial analyses of variance with measured syllable and context syllable as factors.

Results.
The metronome position results are summarized in Table I. It can be seen that in contrast to the general view, the p-center position is not independent of context [2] but significantly influenced by the phonological structure of the second syllable in the same sequence in almost all of the experiments.

Table I:
Metronome position relative to vowel onset levels of significance for the factors S = measured syllable, C = context, I = interaction for the first series of exeriments

	number of experiment 1	2	3	4	5	6
S	.01	.05	.001	.01	.001	.001
C	n.s.	.01	.05	.001	.05	n.s.
I	n.s.	.05	n.s.	n.s.	n.s.	.001

With respect to metronome rate (experiments 1 and 2 vs. 3 and 4) only subject one shows a significant effect on metronome position ($F(1,4) = 9.4$; $p < .05$): with the slower rate (60) of the metronome it is on average 8.5 msec less delayed relative to vowel onset than with a rate of 90 beats per minute.
Experiments 5 and 6 were run to test the influence of the p-center of the metronome signal. Here we used a /vak/-signal with measured metronome delay of 40 msec relative to acoustical syllable onset as metronome. There is a clear effect for both subjects: in this experiment as to be expected the metronome onset is on average 42.1 msec earlier than in experiment 3 for subject one ($F(1,6) = 158.38$; $p < .001$) and 46.2 msec earlier for subject two ($F(1,4) = 121.7$; $p < .001$). For subject one furthermore there is a significant interaction between the effect of metronome signal and measured syllable ($F(6,36) = 3.64$; $p < .01$).
Concerning the predictions of the Marcus formula [2] with respect to the dependency of p-center position on segment durations our results in general are clearly negative. No clear correlation between metronome position and segmental durations could be found for single syllable-context combinations as well as for all measurements of one experiment in single item analysis. A tendency for the predicted depencency can only be found if one considers the pooled data of the single syllable-context combinations. This result seems to suggest that this dependency only holds for phonologically differently composed syllables but not for articulatory variants of phonologically identical syllables.

SECOND SERIES OF EXPERIMENTS

In a second series of experiments we used German monosyllabic verbs varying either the initial consonant, or the vowel length and the syllable final consonance or both. Experiments 7 and 8 were run with verb forms of German "backen" and "packen": /bak/, /bakt/, /bakst/, /pak/, /pakt/, and /pakst/ in all possible syllable-context combinations. In Experiment 9 and 10 we used verb forms of German "spuken" and "spucken": /ʃpu:k/, /ʃpu:kt/, /ʃpu:kst/, /ʃpuk/, /ʃpukt/, and /ʃpukst/ in all possible combinations. Experiment 11 and 12 both were run with subject one. The material consisted of ten repetions of the four possible combinations of "back" and "packst" and of "pack" and "backst", respectively. The metronome rate was set at 50 beats per minute for all experiments.

Results.
The metronome position results are summarized in Table II. Again it can be seen that in almost all experiments there is an influence of context on the position of the p-center.

Table II:
Metronome position relative to vowel onset levels of significance for the factors S = measured syllable, C = context, I = interaction for the second series of exeriments

	number of experiment 7	8	9	10	11	12
S	.001	.001	.05	n.s.	.001	.001
C	n.s.	.001	.001	.05	.01	.05
I	n.s.	.05	n.s.	.05	.001	.001

With respect to the predicted dependency of p-center position on segmental durations our results again are negative: there is no clear correlation to be found in single item analysis.
In experiments 8 and 10 (subject two) parallel to the acoustic recording the glottal opening gesture for the syllable final consonant(s) was registered using an FJ-Photo-Electroglottograph and in experiments 11 and 12 (subject one) orbicularis oris activity was recorded parallel to the acoustic signal using a DISA EMG-Amplifier. For both physiological recordings no correlations with the position of the p-

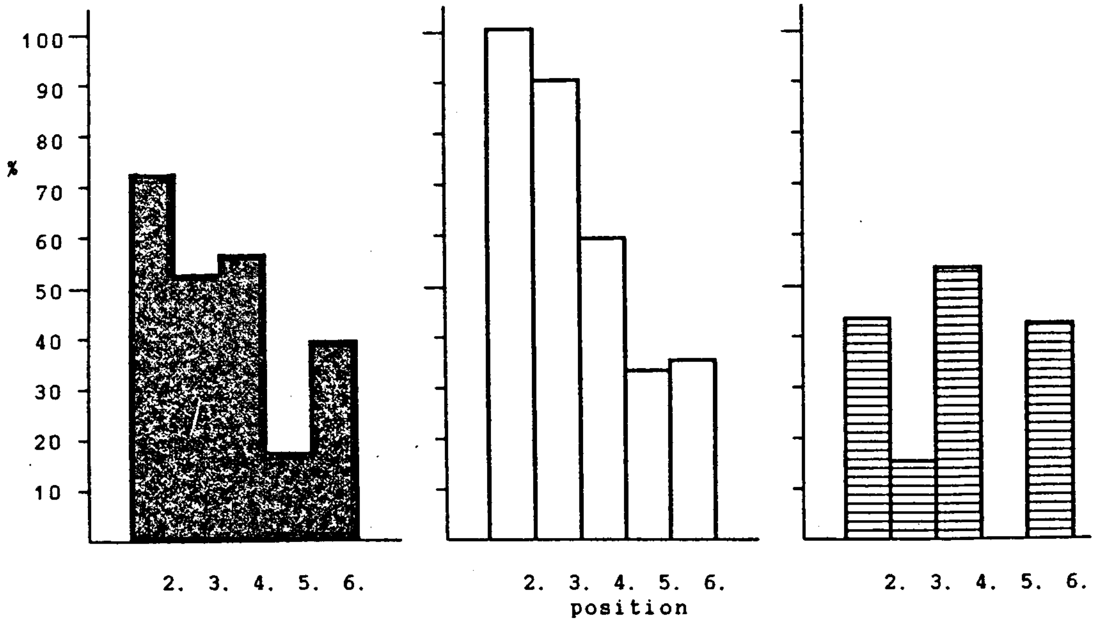


Fig. 1: Delay of the metronome beginning relative to the vowel onset in "packst" in percent total variation (100% = 86.7 msec; 0% = 6.4 msec); left: mean; middle: homogeneous sequences; right: alternating sequences

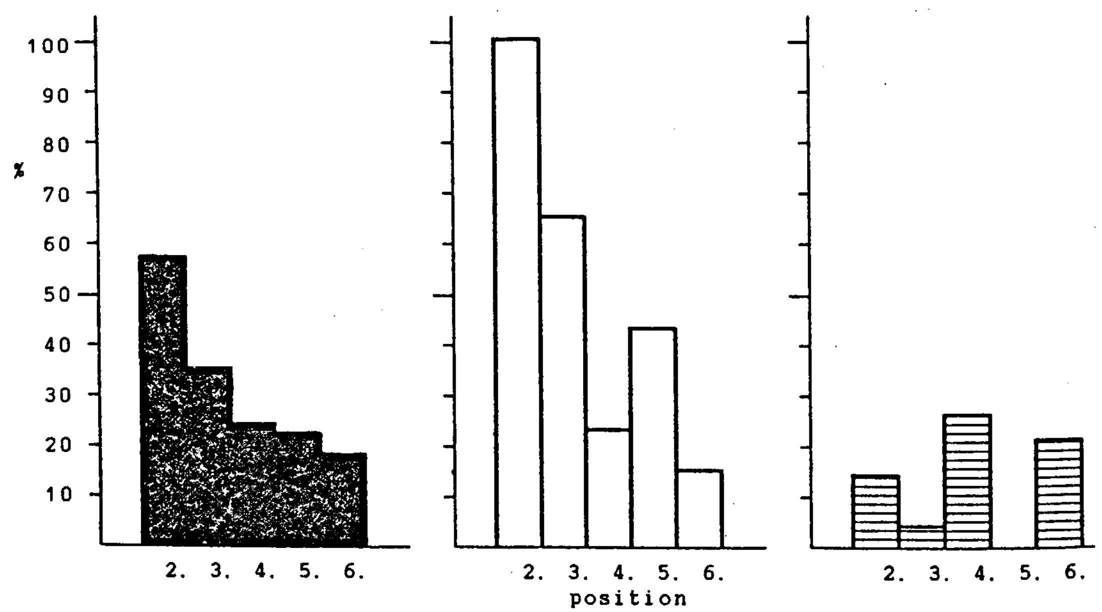


Fig. 2: Delay of the metronome beginning relative to the vowel onset in "backst" in percent total variation (100% = 91. msec; 0% = 7.1 msec); left: mean; middle: homogeneous sequences; right: alternating sequences

center could be found in contrast to the results reported by Tuller & Fowler [3].
The acoustical ddta of experiment 11 and 12 were reanalysed with respect to a possible effect of position of the measured syllable within the sequence on the metronome position. The results are depicted in Figure 1 and 2 for these items of experiment 11 and 12 that show significant effects of position. Two-factorial analyses of variance with as first factor homogeneous vs alternating sequences, and as the other factor position within one sequence showed no effects on metronome position for the items "back" and "pack", but a clear effect for "packst" and "backst": For the item "packst" the metronome position relative to vowel onset is affected by the nature of the sequence ($F(1,89) = 26.2$; $p < .001$), the position within the sequence ($F(4,89) = 8.22$; $p < .001$) and an interaction of both factors ($F(4,89) = 5.69$; $p <$ 001): In homogeneous sequences the metronome delay is longer for the second and third than for the fourth to sixth position, in alternating sequences the second and fourth position (i.e. items of sequences beginning with "back") show longer metronome delays than the third and fifth position (i.e. items of sequences beginning with "packst") and the sixth position differing significantly from the fifth. The simple main effect of the nature of the sequence only is significant for the second, third and fith position: here the metronome delay is longer in the homogeneous sequences.
For the item "backst" the metronome position relative to vowel onset is also affected by the nature of the sequence ($F(1,90) = 17.86$; $p < .001$), the position within the sequence ($F(4,90) = 2.71$; $p < .05$) and an interaction of both factors ($F(4,90) = 4.45$; $p < .01$). The simple main effects are not as pronounced as in "packst": In homogeneous sequences the metronome delay is longer for the second than for the fourth to sixth position and longer for the third than for the fourth and sixth position, but in alternating sequences the tendency paralleling the results of experiment 11 does not reach significance. The simple main effect of the nature of the sequence only is, in parallel to the item "packst", significant for the second, third and fifth position: here the metronome delay is longer in the homogeneous sequences.
This effects seem to result from two general effects: first, metronome delay decreases with position within sequences and second, metronome delay is less for the items with complex syllable final consonance in these alternating sequences beginning with these items.

DISCUSSION

The main result of our experiments shows that for articulatory variants of phonologically identical syllables the p-center position does not show a systematic dependency on segment durations. This dependency can only be seen with regard to the mean values of phonologically differently composed syllables. Moreover, in contrast to the general view, the p-center position is not only an effect of single syllables but significantly influenced by context, i.e. in our experiments by the phonological structure of the second syllable in alternating sequences. It seems therefore that there cannot be a simple acoustical explanation of p-center location based on segmental durations alone.
Nor do our physiological data support a simple articulatory explanation of the p-center phenomenon: in our tests we could not find any correlation between the timing of physiological signals and the position of the metronome onset marked in the acoustical speech signal. The variation of p-center location found in syllables with complex consonantal rhyme alternating with simple CVC syllables, i.e a dependency of p-center position on whether the sequence started with the complex or the simple syllable can only be interpreted on the basis of rather complex articulatory programming.

REFERENCES

[1] Fowler, C.A. 1979, "Perceptual centers" in speech production and perception. Perception & Psychophysics 25, 375-388.

[2] Marcus, S.M. 1981, Acoustic determinants of perceptual center (P-center) location. Perception & Psychophysics 30, 247-256.

[3] Tuller, B., Fowler, C.A. 1980, Some articulatory correlates of perceptual isochrony. Perception & Psychophysics 27, 277-283.

THE ROLE OF AUDITORY CONTROL IN SPEECH MONITORING

KALERIA A.MICHURINA

Dept. of Phonetics
Moscow Institute of Foreign Languages
Moscow, USSR 119034

ABSTRACT

Auditory control is functionally significant in monitoring unskilled speech performance, while skilled performance is rather independent of the auditory sensorium information and is evidently monitored on a higher level than the level of motor or auditory control.

INTRODUCTION

Present-day work does not provide much data concerning self-monitoring of on-going speech. The present study is an attempt to analyse the functional significance of auditory control in speech.
Experimental observations of the effects of delayed auditory feedback point out the critical role of auditory control in monitoring vocal intensity /3/, rhythm and speaking rate /5/ and articulation accuracy /4/.

These and some other findings have been concerned with the importance of audition in monitoring speech automatisms; the experiments were run with adult subjects speaking their native language. However, some investigators of the effects of delayed auditory feedback on non-speech behaviours take the view that the role of auditory control is functionally changeable and depends on the operator's skill. It has been shown that during music performance /1/ and Morse transmission /7/ complex tasks were more greatly disturbed than easier tasks by the delay. The observation is also supported by the experimental data received by J.E.Waters /6/. His subjects from the age of 10 to 18 read outloud under delayed auditory feedback conditions. The older subjects were less affected by the interference and made fewer mistakes than younger subjects. The results may be interpreted in terms of less (younger readers) or more automatized reading skills (older readers). That calls for a closer investigation of the role of auditory control in the period of skill acquisition.

A study of the functional significance of auditory control in monitoring unskilled, semi-skilled and skilled speech performance can provide additional information which speech automatisms obscure. The purpose of the study reported herein is to investigate the problem through experiment.

METHOD

64 college undergraduates (aged 18-24) served as subjects. All of them learned English as a foreign language, with language experience varying from one year (32 subjects, hereafter referred to as beginners) to four years (32 subjects, hereafter referred to as advanced learners). Each of the subjects was required to describe a series of pictures under three conditions:

1) neutral (absence of experimentally induced disturbances),

2) binaural masking (white noise transmitted through earphones),

3) absolute silence (sound-proof earplugs fastened on the head of the subject).

Throughout the session the order of the series within the groups varied and was as follows: 1-2-3, 2-1-3, 2-3-1, 3-2-1,3-1-2. The descriptions were recorded on magnetic tape for later statistical analysis. The samples were analysed for the presence of error performances and the number of self-corrected errors (statistical units totalling 5702).

RESULTS AND DISCUSSION

Our findings support some of the data received under delayed auditory feedback: 1) subjects monitored loudness of speaking (it increased under white noise and decreased under quiet), 2) for both groups speaking rate changed as the result of longer pauses and word or syllable repetitions.
Monitoring of on-going speech proved to be much dependent on the conditions of auditory control. Of the two experimental conditions, absolute silence was more disturbing than white noise.

Beginners were more dramatically affected by auditory disturbances than advanced learners. The interfering influence of their native language (Russian) was felt much stronger under the experimental conditions. First, their articulatory accuracy was seriously impaired. They tended to use Russian substitutes of English sounds, but seemed to be unaware of the fact, as there were no corrections of the errors. Another striking observation is that 7.4% of their sentences under white noise and 5.3% under silence were meaningless in English, but could be traced structurally to analogous sentence patterns in Russian. For example, "She tried to keep the room in tide (tidy)", "They had to live him in their room (let him live)", "They lived in four (there were four of them). Mistakes of the kind were also uncorrected by the speakers. Though the subjects were free to use patterns they knew best, they failed to avoid grammar and lexical errors. The most common classes of grammar errors were articles, tenses, verb forms, and those of lexical - prepositions, choice of words and word-blending. The number of those errors grew in the beginners' speech output from 13% to 27% under white noise and 15% under silence x). In the interpretation of the third figure we have to take into account that the subjects were most reluctant to talk under absolute silence. They tried to escape the situation by making their descriptions very short and by using a limited stock of words and grammar structures.

The subjects' ability to detect and correct errors was significantly affected by the experimental conditions. The percentage of self-corrected errors decreased from 24% in neutral conditions to 18% under white noise and to 12% under quiet.

The overall observation is that the speakers much depended on audition in monitoring their speech performance.

Advanced learners were less affected by the experimental conditions. There was also an increase of errors in their performance, but it was less significant: 8% in neutral conditions, 11% under white noise and 11.2% under absolute silence. Articulatory deviations were few and mainly concerned with full devoicing of final voiced consonants (Russian influence). The mistakes were not corrected either. The subjects' grammar and lexical errors fall into the same classes as those of the beginners, but they are of different nature. Thus, there were no article omissions but articles were often inadequately used; tense mistakes occured not in isolated sentences but inside the sequence-of-tenses paradigm. While beginners used forms, like "feeled", "lucky to got", advanced learners occasionally produced forms that could be taken for blended, e.g. "introduceded", "they got marriaged". The errors of word choice were more numerous in the speech of advanced learners, which is probably due to a higher level of lexical programming. For example: "She was over hair (head) and ears in work", "Ted and Ann were a newly-made (married) couple". Enigmatically, their number decreased under experimental conditions.

Unlike beginners, the advanced learners were not at all reluctant to talk under absolute silence. They even tried to make their descriptions more "beautiful" which resulted in quite a number of bookish, unnatural expressions. For example, "He came into their view with a girl", "He changed his figure into a stout one", "His thought gave result".

Error-correction of advanced learners was less effective under experimental conditions. The corresponding percentage was 17% for neutral conditions, 13.8% under white noise and 12% under absolute silence.

On the basis of data received we assumed that semi-skilled speech performance is rather independent of the auditory sensory alterations.

In an attempt to support the assumption we asked 10 college teachers with a perfect command of English to give their descriptions of the pictures under white noise and absolute silence. Neither of the conditions had any consistent effect upon their speech performance. The only alterations were longer pauses and an increased number of word repetitions. The subjects were fully able to monitor their performance without the benefit of auditory control. In fact, there were only two slips, and these were immediately corrected.

x) These and other figures are statistically significant.

CONCLUSIONS

The results of the study show that the functional significance of auditory control in monitoring unskilled, semi-skilled and skilled speech performance is different.

It is the greatest in monitoring unskilled performance, significantly reduced in monitoring semi-skilled performance and is minimal in monitoring skilled speech performance.

This seems to be well related to N.Bernstein's idea of a multi-level natur of control /2/. In reference to speech it can mean that unskilled speech performance is monitored on the basis of current sensorium information, while skilled performance is monitored on a higher level (hypothetically, sense-level) which is independent of the afferent information received by the ear.

In the period of deranged speech automatisms restoration auditory control may even become a hindrance. We observed two brain-damaged patients during their rehabilitation period. They showed much better speech performance when talking with the ear plugs than when they were aware of the acoustic effect of their performance.

This may turn out true in cases of stutters.

Altering audition may serve as an effective means of checking the degree of speech habit formation. When formed, they remain intact.

The way we see it, the results of the findings may be applied in language teaching and logopedics.

References

1. Ansell, S.D. Delayed auditory feedback and human skill. "Dissertation Abstracts", v.26, No.5554. USA, 1965.
2. Bernstein, N.A. The coordination and regulation of movements, Moscow: Medgiz 1947.
3. Black, J.W. The effect of delayed side-tone upon vocal rate and intensity "Journal of speech and hearing disorders". 16, 1951.
4. Fairbanks, G., Guttman, N. Effects of delayed auditory feedback upon articulation. "Journal of speech and hearing disorders", v.I. No.I, 1958.
5. Kozhevnikov, V.A. and Chistovich, L.A. (eds.). Speech: Articulation and Perception, Moscow: Nauka, 1965.
6. Waters, J.E. A theoretical and developmental investigation of delayed speech feedback. "Genetic psychology monographs". 78 (I), 1968.
7. Yates, A.J. Effects of delayed auditory feedback on Morse transmission by skilled operators. "Journal of experimental psychology". 69 (5), 1965.

МИКРОСЕГМЕНТЫ КАК ОСНОВНЫЕ ЭЛЕМЕНТЫ ПЕРВИЧНОЙ СЕГМЕНТАЦИИ РЕЧЕВЫХ СИГНАЛОВ

В.Г. РУДАКОВ

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР
АН СССР, МОСКВА

В.Н. ТРУНИН-ДОНСКОЙ

ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР
АН СССР, МОСКВА

В докладе рассматривается возможность использования для целей первичной сегментации речевых сигналов микросегментов, определяемых в виде совокупности локальных длительностей. И микросегменты и локальные длительности определяются непосредственно из анализа формы речевой волны во временной области. Использование микросегментов позволяет примерно в 2 раза сократить исходную длительность сигнала для последующего анализа на фонемном уровне, а также указать некоторые параметры фонем.

Известно /I,2/, что вся информация о речевом сигнале содержится в его временной функции $P(t)$, отражающей зависимость звукового давления P на некотором расстоянии от говорящего. Успешному решению ряда проблем анализа речевых сигналов способствует правильное проведение процесса их сегментации/3/. С точки зрения достижения максимальной информативности результатов анализа сегментация должна осуществляться адаптивным способом к последовательным во времени звуковым явлениям /4/. Для первичной сегментации речевой сигнал представляют в виде последовательности вокализованных и невокализованных сегментов/I/. В /3/ в качестве основных элементов первичной сегментации речевого сигнала предложены микрофонемы - участки сигнала на протяжении периода основного тона. Преимуществом такого способа является то, что микрофонемы не связаны со строго постоянным интервалом в I0...20 мс, а также большая уместность микрофонемных и фонемных характеристик, относящихся к одному и тому же диктору и классу звуков речи. Недостатком описываемого способа сегментации является использование для анализа сигнала быстрого преобразования Фурье, обусловленного допущением о квазипериодичности и стационарности сигнала на всём протяжении T_0. Такое допущение к речевому сигналу не совсем справедливо /4/. В /4/ речевой сигнал предложено представлять в виде сложной кривой, а для её анализа - метод разложения сложных кривых на компоненты. Этот метод справедлив для анализа как вокализованных, так и невокализованных сегментов, но он, как и спектральный метод,не предполагает непосредственного анализа формы речевой волны.

Для выявления некоторых параметров, характеризующих форму речевой волны рассмотрим на рис. I фрагмент осциллограммы преобразованного в электрический сигнал $U(t)$ изменения звукового давления $P(t)$ на интервале $[t_0, t_N]$.

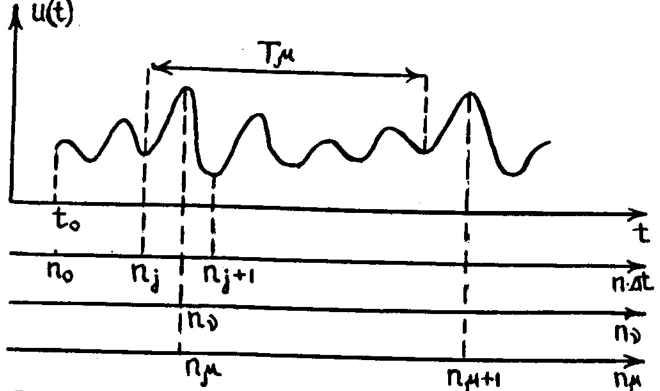


Рис. I. Фрагмент осциллограммы преобразованного в электрический сигнал $U(t)$ речевого сигнала $P(t)$.

Произведем замену непрерывного времени t на дискретное $n \cdot \Delta t$. Принимая $\Delta t = const$, получим зависимость $U(n)$, аргументом которой является номер дискреты $n \in [n_0, n_N]$.

Очевидный колебательный характер функции $U(n)$ можно описать с помощью следующих параметров. На первом уровне описания используются временные интервалы между локальными экстремумами. В номере дискреты n имеет место локальный минимум, если выполнено условие

$$[u(n-1)>u(n)]\wedge[u(n)<u(n+1)],$$

$$u(i)>0,\ i=n-1,n,n+1. \qquad (I)$$

Для обозначения номеров дискрет, в которых выполняется условие (I) введём индекс j. Смежные локальные минимумы $U(n_j)$ и $U(n_{j+1})$ определяют j-е длительности

$$\tau_j=(n_{j+1}-n_j)\cdot\Delta t=\Delta n_j\cdot\Delta t \qquad (2)$$

Локальные максимумы определяются в пределах изменения τ_j при выполнении условия $[u(n-1)<u(n)]\wedge[u(n)>u(n+1)]$,

$$u(i)>0,\ i=n-1,n,n+1. \qquad (3)$$

Номера дискрет, удовлетворяющие условию (3) обозначим с помощью индекса ν, тогда локальные максимумы будут иметь обозначение $U^{(j)}(n_\nu)$, где индекс j указывает на их принадлежность к соответствующей τ_j. Длительности τ_ν находятся аналогично выражению (2). Поскольку длительности τ_j и τ_ν определяются соответственно смежными минимумами и максимумами с перекрытием $0{,}5\tau_j$ $(0{,}5\tau_\nu)$, то их наложение может быть использовано как для исключения помех, так и для выявления дополнительных сведений о тонкой структуре сигнала.

Таким образом на первом уровне анализа речевой сигнал представляется последовательностью чисел, характеризующих локальные экстремумы $U(n_j)$, $n_j \in (n_0, n_N)$, $U^{(j)}(n_\nu)$, $n_\nu \in (n_0, n_N)$ и длительности τ_j и τ_ν.

На втором уровне анализа производится выделение значимых экстремумов из локальных. С целью наибольшего учёта динамики функции $P(t)$ необходимо использовать такие однотипные экстремумы, дисперсия которых наибольшая. Проведенный анализ показал, что этому условию удовлетворяют локальные максимумы, поскольку их дисперсия примерно в 4 раза превышает дисперсию минимумов. Значимый максимум определяется из анализа условия

$$[U^{(j-1)}(n_{\nu-1})<U^{(j)}(n_\nu)]\wedge[U^{(j)}(n_\nu)>U^{(j+1)}(n_{\nu+1})]. \qquad (4)$$

Введём индекс μ для переобозначения таких номеров дискрет n_ν, для которых условие (4) выполняется. Очевидно, что значимый максимум $U^{(j)}(n_\mu)$ всегда совпадает с соответствующим локальным максимумом $U^{(j)}(n_\nu)$.

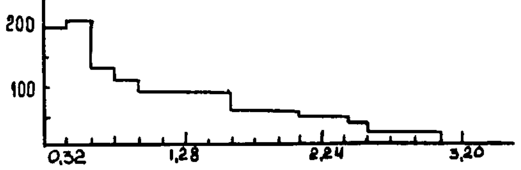
Это позволяет указать временной интервал между смежными значениями максимумов $U^{(j)}(n_\mu)$ и $U^{(j+k)}(n_{\mu+1})$, где k - количество локальных максимумов между номерами дискрет n_μ и $n_{\mu+1}$. Обозначим его через T_μ и определим с помощью выражения

$$T_\mu=(n_{\mu+1}-n_\mu)\cdot\Delta t=\sum_{j=1}^{k+1}\tau_j. \qquad (5)$$

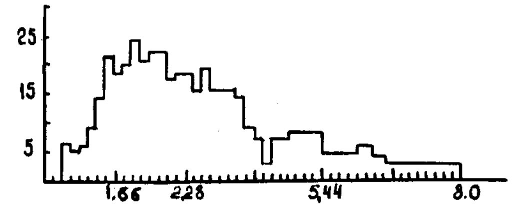
Введём рабочую гипотезу о том, что форма речевой волны в первом приближении может быть охарактеризована параметрами $U(n_j)$, $U^{(j)}(n_\nu)$, τ_j, τ_ν на интервале T_μ, $n_\nu \in [n_j, n_{j+k}]$, $k \in (n_\mu, n_{\mu+1})$.

Для экспериментальной проверки возможности описания формы речевой волны с помощью введённых параметров был использован словарь из 27 слов: ноль, нуль, один, два, три,четыре, пять, шесть, семь, восемь,девять, действие,сложить, вычесть, умножить, величина, точка, цифра, синус, косинус, тангенс, котангенс, слушай, начало, конец, число, целое. Этот словарь по-слогно 2-мя мужчинами и женщиной разговорным стилем в помещении машинного зала с уровнем шумов 65 дБ по телефонному каналу с полосой частот 3, I25 кГц, передавался на вход I0-разрядного преобразователя аналог-цифра. В соответствии с указанной полосой частота дискретизации принята 6,25 кГц, что соответствует Δt = 160 мкс.

По результатам обработки 8I слова из указанного словаря на рис. 2 приведены гистограммы для значений длительности $\tau_{j(\nu)}$ и T_μ.



Локальные длительности $\tau_{j(\nu)}$, мс



Длительности T_μ, мс

Рис. 2. Гистограммы длительности $\tau_{j(\nu)}$ и T_μ. Данные получены в результате обработки 8I слова, произнесённых 3-мя дикторами по 27 слов каждый.

Анализ приведенных гистограмм показывает, что локальные длительности $\tau_{j(\nu)}$ занимают диапазон от 320 мкс до 3 мс с максимумом в районе 320 мкс, что соответствует частоте 3,I25 кГц, то есть верхнему значению спектра сигнала. Диапазон значений T_μ находится в пределах от 640 мкс до 8 мс с максимумом на T_μ = 2мс или 500 Гц. Поскольку эти области существенно перекрываются, то они могут быть использованы в ограниченных целях, например, для определения высоты голоса по положению максимума гистограммы длительности T_μ для одного диктора.

Анализ чередований T_μ на протяжении отдельных слов показал, что они обладают определёнными регулярностями. В первом приближении эти регулярности могут быть описаны с помощью семи правил (ПI,..П7), которые удовлетворяют следующим выражениям:

$$\text{П1}, |T_\mu - T_{\mu+1}| \leq 160 \text{ мкс} \qquad (6)$$

$$\text{П2}, |(T_\mu + T_{\mu+1}) - (T_{\mu+2} + T_{\mu+3})| \leq 160 \text{ мкс} \qquad (7)$$

$$\text{П3}, |(T_\mu + T_{\mu+2}) - (T_{\mu+1} + T_{\mu+3})| \leq 160 \text{ мкс} \qquad (8)$$

$$\text{П4}, |T_\mu - (T_{\mu+1} + T_{\mu+2})| \leq 160 \text{ мкс} \qquad (9)$$

$$\text{П5}, |(T_\mu + T_{\mu+1}) - T_{\mu+2}| \leq 160 \text{ мкс} \qquad (I0)$$

$$\text{П6}, |(T_\mu + T_{\mu+2}) - T_{\mu+1}| \leq 160 \text{ мкс} \qquad (II)$$

$$\text{П7}, |(T_\mu + T_{\mu+1} + T_{\mu+2}) - (T_{\mu+3} + T_{\mu+4} + T_{\mu+5})| \leq \qquad (I2)$$

≤ 160 мкс

где 160 мкс = Δt.

Длительности T_μ, удовлетворяющие в своей последовательности соответствущему правилу, объединяются в группы. В речевых сигналах на этом этапе наблюдаются чередования как одинаковых, так и разных групп, которые образуют макрогруппы. Между макрогруппами, а иногда и между группами, встречаются длительности, которые не удовлетворяют приведенным в выражениях (6)...(12) правилам объединений. Эти длительности не используются для анализа на них распределений $\tau_j(\nu)$. Они могут быть учтены лишь при решении вопроса о наличии либо паузы, либо помехи в слове.

Для иллюстрации отмеченных этапов анализа чередования длительностей T_μ на рис. 3 приведена гистограмма объединения T_μ в группы в соответствии с правилами П1,...,П7 в слове"число", диктор мужчина. Из рисунка 3 следует, что каждому правилу объединения T_μ соответствует определённое число групп.

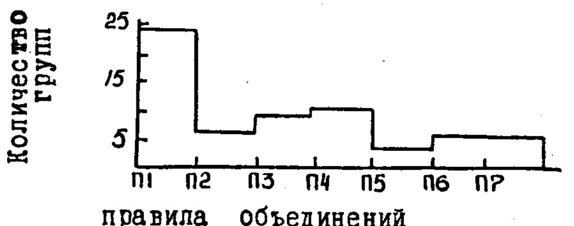


Рис. 3. Гистограмма объединений длительностей в группы с использованием правил П1,...,П7 в слове "число".

Анализ гистограмм слов приведенного выше словаря показал, что для целей сегментации следует выбирать либо такие группы, у которых количество T_μ не менее 7, либо макрогруппы с числом однотипных групп не менее 4 и количеством T_μ в них более 4. Существующую в слове"число"последовательность обозначим в виде следующих групп: П7.1; П7.2; П1.3; П7.4; П7.5; П7.6; П2.7; П2.8; П7.9, где вторая цифра после номера правила объединения обозначает порядок следования групп.

Анализ этих последовательностей показал, что группы П1.3, П2.7 и П2.8 соответствуют вокализованным сегментам, так как у них нет дробления длительностей T_μ. В группе П1.3 они находятся в диапазоне от 1,32 до 2,4 мс,в группе П2.7 - от 2,72 до 3,52 мс, в группе П2.8 - от 2,88 до 3,52 мс. Второй характеристикой этих сегментов является распределение $\tau_j(\nu)$. В группе П1.3 длительности $\tau_j(\nu)$ в основном находятся в области от 0,32 до 0,96 мс, в П2.7 - от 1,12 мс до 2,56 мс, а в П2.8 - от 1,44 до 2,08 мс. Распределения T_μ и $\tau_j(\nu)$ на них в остальных группах характеризуют невокализованные сегменты. Ориентировочно приведенные группы могут быть соотнесены с фонемами: П7.1, П7.2 - "Ч"; П1.3 - "И"; П7.4, П7.5 - "С"; П7.6 - "Л"; П2.7, П2.8 - "О". Следовательно, по распределениям такого типа можно приблизительно производить сегментацию речевых сигналов на вокализованные и невокализованные участки, а также выносить определённые суждения о фонемных характеристиках выделенных сегментов.

В заключение рассмотрим таблицу;в таблицу сведены соотношения длительностей приведенных групп с длительностью слова.

Из таблицы следует, что общая длительность слова "число" составляет 623,6 мс, а длительность групп - 181 мс, что соответствует 29% от длительности слова. Кроме того, длительности групп: П7.1, П7.2,П7.4, П7.5, П7.6 и П7.9 в среднем совпадают с общепринятым окном анализа в 10...20 мс.

Таблица. Представление длительности слова "число" через длительности групп и интервалов между ними

	Длительность, мс								
Группы	П7.1	П7.2	П1.3	П7.4	П7.5	П7.6	П2.7	П2.8	П7.9
Длительности групп	12,9	8,3	28	15,8	18	16,6	42,8	32,8	13,9
Интервалы между группами	70	43,6	69,1	43,3	0	42,5	0	166	

Всё это позволяет сделать вывод о целесообразности использования в качестве основных элементов первичной сегментации речевые сигналы с длительностями T_μ, которые предлагается называть микросегментами.

ЛИТЕРАТУРА:

1. Фант Г. Акустическая теория речеобразования. Пер. с англ. под ред. В.С. Григорьева. - М.: "Наука", 1964, 283 с.
2. Линдсей П.,Норман Д. Переработка информации у человека. Пер. с англ. под ред. А.Р. Лурия. - М.: "Мир", 1974, 550 с.
3. Джерниковский А. Микрофонемы как основные сегменты первичной сегментации речевого сигнала. - Автоматическое обнаружение микрофонем. В Трудах IУ Международной объединенной конференции по искусственному интеллекту. Тбилиси,1975, том 5,с.68 - 82.
4. Соломатин В.Ф. Метод разложения сложных кривых на компоненты. Деп. ВИНИТИ, № 4967-81,- М.; 1981, 15 с.

АВТОМАТИЧЕСКАЯ СЕГМЕНТАЦИЯ СОЧЕТАНИЙ ЗВУКОВ (ДИАД)

ПЁТР ДОМАГАЛА

Лаборатория Акустической Фонетики
Институт Основных Проблем Техники ПАН
Познань

РЕЗЮМЕ

Представлена система сегментации сигнала речи на основании изменений распределений энергии в соседних спектрах. Применён алгоритм, основанный на нескольких логических зависимостях, что требует низких вычислительных мощностей ЭВМ. Система применена для сегментации 444 польских диад, произнесённых 8 дикторами. Получено в среднем 90% правильных сегментаций.

ВВЕДЕНИЕ

Исследования, касающиеся автоматической сегментации сигнала речи (АСР), обычно представляют собой предварительный и существенный этап сложного процесса автоматического распознавания речи (АРР). В области проблематики анализа речевых сигналов человеком или машиной понятие сегмента применяется, вообще говоря, для определения временного фрагмента единицы (обычно лингвистической), являющейся элементом во множестве знаков, подвергающихся распознаванию. Таким множеством может быть слово или фраза. В таком случае сегментом может быть звук, сочетание двух звуков (диада), либо слог. В случае модели распознавания биологической системой сегмент не может выбираться произвольно, а должен соответствовать фактическим физиологическим, неврологическим и психологическим процессам, имеющим место в восприятии. При автоматическом распознавании не обязательно копирование этих процессов, и поэтому сегментом может быть фрагмент сигнала, который не совпадает ни с элементом восприятия, ни с лингвистическим элементом, разве что предполагается, что в рамках проблематики бионики специально конструируются системы, копирующие биофизические функции. Несмотря на то, что в настоящее время существует много моделей и теорий восприятия речи /3/, часто противоречащих одна другой, имеется значительное единство мнений, касающееся того, что в нормальном процессе восприятия речи человеком в рамках таких более крупных лингвистических единиц как предложение (высказывание) или слово, происходит выделение по крайней мере некоторых сегментов, длительностью приближающихся к звуку. В то же время не ясно, на каком уровне восприятия происходит звуковая сегментация – слуховом, фонетическом или высшем. Сегмент является элементом конечного множества значительно меньшей численности, чем численность множества элементов высшего ряда – для данного законченного словаря число слов гораздо больше числа слогов, которое, в свою очередь, больше числа фонем.
В настоящей работе понятие сегмента отождествляется с фонетической единицей, понимаемой как моносегментный звук, либо с сегментом полисегментного звука. Представлена АСР, опирающаяся на анализ изменений распределения энергии в спектре, а также результаты сегментации чаще всего встречающихся в польском языке диад.

ОПИСАНИЕ МЕТОДА

Для реализации метода была использована аналого-цифровая система, в состав которой входят: 60-канальный анализатор спектра, интерфейс, соединяющий аналоговый источник сигнала с микро-ЭВМ и микро-ЭВМ МЕРА-303. Аналоговый анализатор спектра имеет 43 полосы постоянной ширины, составляющей 80 гц, покрывающие область частот от 120 до 3560гц, а также 17 полос с шириной, зависящей линейно от среднегеометрической частоты и покрывающей диапазон от 3560 гц до 7000 гц. Выходы отдельных каналов циклически подключаются к общему выходу. Полученная цифровая спектрограмма сигнала речи после усреднения остаётся в памяти ЭВМ в виде таблицы с координатами времени и частоты /1/. Для каждой пары очередных спектров $k-1$ и k создан N – элементный ряд (N – число каналов) с элементами $r_{ik} = a_{ik-1} - a_{ik}$, где $i = 1,\ldots,N$ обозначает номер полосы, K – очередной квант времени. Ряд r_{ik} был поделён на $c(k)$ составных рядов при применении критерия согласованности знака и достаточно высоком абсолютном значении, то есть направления и скорости изменения уровня. Через $z(k)$ обозначен знак элементов последнего составного ряда, принимая 0 для положительных величин и 1 для отрицательных. Принято, что граница между сегментами будет обозначена в следующих случаях:

1) $c(k)= 1 \wedge [c(k+1) \neq 1 \vee [c(k+1) = 1 \wedge z(k) \neq z(k+1)]]$

2) $c(k) = c(k+i) = 1 \wedge z(k) = z(k+i) \wedge [c(k+n) \neq 1 \vee [c(k+n) = 1 \wedge z(k) = z(k+n)]]$
где $i=1, 2,\ldots, n-1$

3) $c(k)= 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \; c(k+1) \neq 2$

4) $c(k)= 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge z(k) \neq z(k+n) \wedge c(k+n) = 2$ где $i=0, 1, \ldots, n-1$ а $n \geqslant 1$

5) $c(k)= 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge c(k+n) \neq 1 \wedge c(k+n) \neq 2$
где $i=0, 1, \ldots, n-1$ а $n \geqslant 2$.

Вышеуказанные логические зависимости были введены в систему МЕРА 303. Для подготовки экспериментального материала были использованы данные, полученные Яссемом, Лобач в их работе, касающейся фонотактики польского языка /2/. Опубликованный в этой работе список чаще всего встречающихся в польском языке диад в 94% охватывал анализируемую там выборку численностью 10^5. Этот список был использован с исключением диад типа: "#F", "F#" и "F_jF_j" (# обозначает паузу, F – какую-либо фонему, а "F_jF_j" обозначает диаду, состоящую из одинаковых фонем). Это обозначало сокращение списка до 444 пар различающихся между собой фонем.

Для каждой диады было образовано искусственное слово (логатом), содержащее две её реализации и отвечающее принципам фонотактики польского языка. Созданные логатомы были сгруппированы в четыре списка в очерёдности, соответствующей порядку появления диад в списке частотности. Отдельные списки содержали 112, 109, 105 и 118 логатомов каждый. Каждый логатом был произнесён по 3 раза восьмью дикторами (5 мужских голосов и 3 женских голоса) и записан на магнитную ленту. Представленный выше материал был подвержен автоматической сегментации. В качестве порогового значения скорости изменения уровней сигнала в отдельных каналах спектрального анализатора были приняты примерно 30 дБ/23 мсек (23 мсек – это временное расстояние между соседними спектрами). Это значение было постоянным для всех голосов. Была проанализирована частота и проведена автоматическая сегментация каждого воспроизведённого с магнитной ленты логатома. Спустя примерно 1,5 сек на экране монитора появлялась спектрограмма высказывания с обозначенными границами. Изображение оставалось неподвижным в течение примерно 5 секунд, т.е. до времени появления следующей спектрограммы. В это время следовало найти те фрагменты спектрограммы, которые относились к диадам, являющимся основой конструкции логатома. Если положение автоматически определённой границы совпадало с серединой переходного участка между двумя фонемами с точностью расстояния между двумя соседними спектрами (23 мсек), то сегментация считалась правильной. Это расстояние (в два раза большее, чем применяемое в других методах) превышает длительность самых коротких артикуляционных явлений. Условием положительной оценки сегментации диад, содержащих полисегментальную фонему, было выделение основного сегмента, например, смычки в согласных смычных звуках. Каждый логатом произносился по три раза одним и тем же диктором и, следовательно, каждая диада была произнесена 6 раз. Записывались результаты 5 первых повторений.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В Табл. 1 представлены средние значения эффективности АСР по отдельным голосам, для мужских голосов и женских (м,ж), а также для целой группы (мж). Значения, находящиеся в колонках, обозначенных "а", являются средними арифметическими, полученными в эксперименте, а данные в колонках, обозначенных "б", являются средними, учитывающими распределение частоты встречаемости анализированных диад в польском языке. В Табл.2 представлены средние значения и дисперсии отсутствующих сегментаций, приходящихся на одну диаду. Диады, чаще встречающиеся в польском языке (Список 1), легче поддаются сегментации, чем диады, встречающиеся реже (Список 4). Из рис. 1 вытекает, что диады, сегментируемые с эффективностью не менее 80% (в среднем 1 ошибка на голос) составляли 75% всех диад. Принимая во внимание частотность, устанавливаем, что величина эта возрастает на 5%. Для всего материала (444 диады×8 дикторов×5 повторений) получено 90% эффективности сегментации. В Табл. 3 представлено распределение типов диад из анализируемого материала вместе с данными, касающимися процента отсутствующих сегментаций. (C обозначает согласный, V – гласный, $\breve{V}$ – не образующие слогов j, ω). Самой большой податливостью на сегментацию отличаются диады типа CV и VC. Причины низкого уровня правильной сегментации ниже 60% для некоторых диад следующие:

- малая контрастность /eo, nu, un, ʃu, um, oa, uω, fʃ, oω, ao, ea, t̂j, xf, ji, ωu, ij, fs/;
- низкий энергетический уровень обеих фонем /ωt, nt͡s, vu, nm, ug, nd͡z, gr, mv, nt, ng, uv, ωt, mp, nd, mb/.

Первая причина имеет объективный характер, вторая вытекает из специфики метода.

Представленный и применённый метод АСР характеризуется относительно высоким уровнем эффективности сегментации при использовании минимальных вычислительных мощностей, что гарантирует работу в условиях, приближённых к реальному времени. Метод можно использовать в качестве предварительного этапа процедур параметризации и распознавания речи. Полученные результаты для чаще всего встречающихся в польском языке диад можно использовать для создания искусственных языков для целей коммуникации человек – машина.

Голос	Список 1		Список 2		Список 3		Список 4		Вместе	
	а	б	а	б	а	б	а	б	а	б
1м	90,8	90,4	90,6	90,0	88,3	89,6	89,3	88,6	89,8	90,1
2м	94,0	95,8	92,1	92,4	87,4	86,5	84,6	84,2	89,5	93,6
3м	83,6	82,8	88,4	88,4	80,4	79,2	81,7	82,4	83,5	83,5
4м	96,6	96,0	94,1	94,4	90,6	90,0	93,9	94,0	93,8	95,0
5м	96,1	96,5	81,5	82,1	89,3	89,4	80,2	79,8	86,7	91,9
6ж	89,3	88,6	87,2	87,2	85,7	84,1	82,8	83,5	86,2	87,7
7ж	91,1	90,2	83,7	83,6	70,0	71,1	84,0	83,7	82,4	86,6
8ж	92,1	92,1	87,7	87,7	80,0	79,8	77,4	78,9	84,4	89,2
м	92,2	92,3	89,3	89,5	87,2	86,9	85,9	85,8	88,6	90,8
ж	90,8	90,3	86,3	86,2	78,6	78,4	81,4	82,0	84,3	87,8
мж	91,7	91,5	88,2	88,2	84,0	83,7	84,2	84,4	87,0	89,7

Табл. 1 Эффективность автоматической сегментации в %

Список	Число диад	Число отсутствующих сегментаций	Среднее значение	Дисперсия
1	112	382	3,41	20,33
2	109	516	4.61	47.19
3	105	671	6.39	47.23
4	118	753	6.38	52.64

Табл. 2 Отсутствующие сегментации в списках логатомов

Тип диады	CV	VC	CC	VV	V̆C	CV̆	V̆V	VV̆	V̆V̆
Число диад	118	148	118	7	13	18	11	11	0
% отсутств. сегментаций	6.2	9.9	17.1	44.0	23.3	16.3	29.8	34.1	-

Табл. 3 Сегментация типов диад

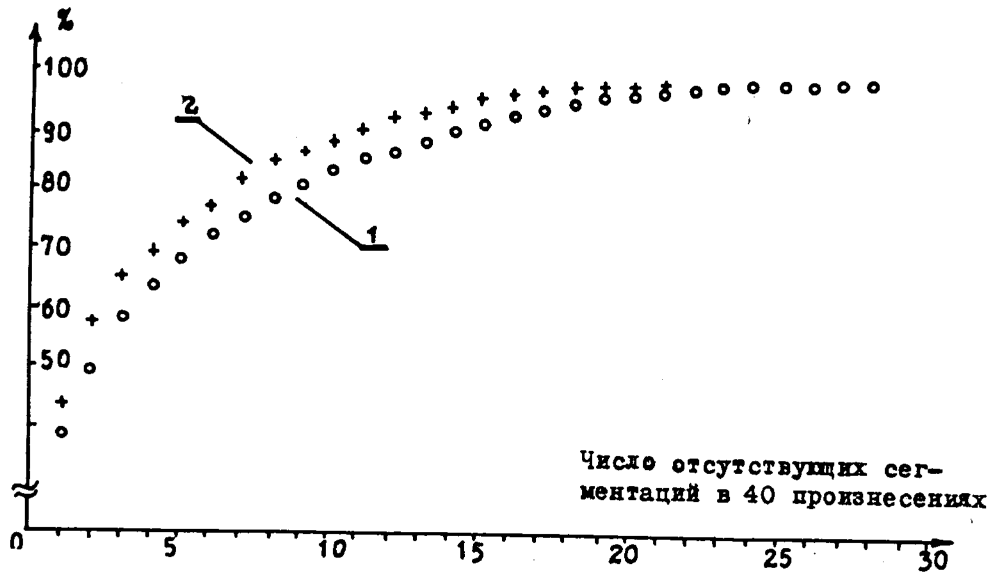


Рис. 1 Функция распределения отсутствующих сегментаций, 1 - без принятия во внимание и 2 - с принятием во внимание частотности диад

ЛИТЕРАТУРА

/1/ Домагала, П., Автоматизация процесса сегментации сигнала речи в аналого-цифровой системе, Работы Ин-та Основных Проблем Техники, №5/1984, Варшава.

/2/ Яссем, В., Лобач, П., Фонотактический анализ польского текста, Работы Ин-та Осн. Пробл. Техн., №63/1971, Варшава.

/3/ Лобач, П., Фонетико-лексикальные взаимодействия в восприятии речи, Изд. Ун-та А. Мицкевича, Познань, 1985.

ONE OF THE METHODS OF AUTOMATIC SYLLABLE SEGMENTATION FOR CONNECTED SPEECH

RODMONGA POTAPOVA

Dept. of Experimental Phonetics, Institute of Foreign Languages, Moscow, USSR 119034

ABSTRACT

The present report proposes a method of automatic speech segmentation using syllable templates based on soft-hardware method of speech analysis to cope with several difficulties, i.e. indistinct syllable boundaries; absence of data on the amount of phonemes in a syllable, localization of phonemes on temporal axis; relative complexity of processing. The report describes one of the possible approaches to continuous speech segmentation, which is of great importance in solving tasks of automatic recognition and understanding of a spoken message in the process of "man-to-computer" communication using a natural language and spoken speech as the basis.

INTRODUCTION

Method based on syllable templates is widely used in automatic speech recognition systems. At present three approaches to automatic speech recognition using syllable templates are known:
- a) input speech is segmented into syllable-sized units which are matched against stored syllable templates;
- b) words synthesied from syllable-sized units are matched against input words;
- c) input speech signal is analyzed, segmented into soundlike (or smaller) segments with subsequent forming of syllable units.

From literature it is known that in the first type of methods the difficulty is that they are liable to segmentation errors, while the difficulty in the second and the third approaches is an increase in the complexity of processing. Though the method using syllable templates is rather effective because it takes into account most of coarticulation phonemes models without considerable extention of memory and increase of processing rate, it is limited now, first of all, by the way of presentation of input material (separately pronounced words) and limited number of speakers [2].
Complexity of speech recognition as the result of increase of a number of speakers and extention of lexicon, clearly shows the advantage of the syllable segmentation method based on changes of feature parameters of speech wave [4].
It is proposed that the method can integrate with the method using syllable templates because, for example, in the Russian language there are comparatively small amount of main syllable types ($n \approx 200$). It is possible to form plausible hypotheses on phoneme structures of a segment using data of probable occurence of these syllables and of phonetic correlates of distinctive phoneme features forming these syllables [5, p.98].

PROPOSED METHOD

Well-known principles of syllable recognition of speech are based either on analysis of average signal energy $E_{(t)}$ (the envelope) and determination of minimum and maximum of the envelope, intervals between minima of signal envelope being taken for syllable boundaries and maxima of the envelope being located on the nucleus of a syllabic vowel; or on the analysis of the wave itself by segmenting it according to maxima and minima with subsequent forming of syllable units based on typical properties of segments, mainly, of an energy character [3].
In the method of syllable segmentation there are several difficulties:
- indistinct boundaries between syllables;
- absence of data on the amount of phonemes in a syllable and on the localization of these phonemes in time;
- heuristic approach in forming syllable units from segments;
- relative complexity of processing;
- insufficient self-descriptiveness of parameters which have very often nothing to do with parameters of vocal apparatus with subsequent false maxima and minima which are the result of energy changes in bands depending on peculiarities of vocal apparatus more than on peculiarities of some parameters.

The present method of automatic speech recognition using syllable templates is free from these defects. It is based on

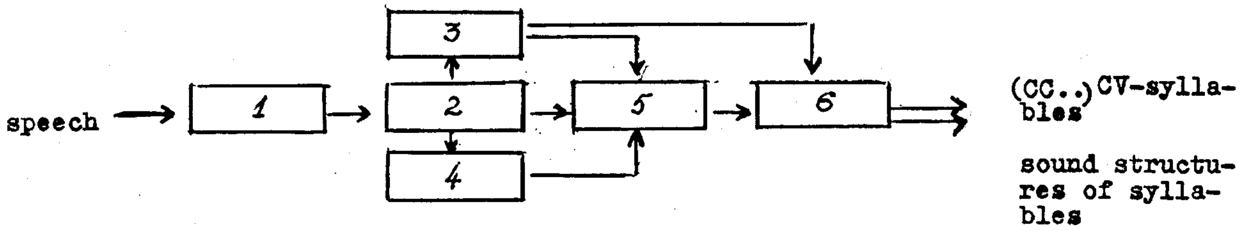


Fig.1. Block diagram of automatic speech segmentating process using syllabic units

soft-hardware method of analysis of speech sounds. This is achieved:
-a) by segmentation of several parameters from a speech sound, (formant frequency, intensity averaged by analysis interval, frequency averaged by transition of signal through zero, pitch, etc) by analog or digital processing;
-b) by segmentation of a sequence of parameters correspondent to speech phrase;
-c) by analysis of a sequence of segment;
-d) by summery up derivatives of parameters normalyzed and averaged by all parameters in time;
-e) by obtaining segmental function $S_{(t_j)}$;
-f) by location of extreme maxima of this function in a signal which are characterized by intensity decrease;
-g) by taking the extreme maxima for the boundaries of open syllables and extreme maxima $S_{(t_j)}$ between boundaries of syllables for the boundaries of sounds in every syllable.
The value of the function at the moment t_j is:

$$S_{(t_j)} = \sum_{i=1}^{n} \frac{k_i \Delta A_{i,t_j}}{\bar{\Delta} A_{i,t_j}}$$

A_i is the i-th feature, where k_i is the weight value of a parameter, n is the number of speech parameters utilized.
Exact boundaries of (CC...)CV-syllables are determined by the highest peaks of S_{t_j} in time, that are characterized by intensity decrease.
Peaks of $S_{(t_j)}$ inside the syllable boundaries determine boundaries between sounds.
The organization block diagram is illustrated as follows (Fig. 1).
-Signal processed is put into segmenter of speech parameters (block 1) as described in [1].
-From input register sequence of parameters correspondent to input utterance is transferred to storage (block 2).
-Stored data are processed by blocks 3 and 4.
-Block 2 using the sequence of parameters finds and stores location on temporal axis and quantity of maxima of $S_{(t_j)}$.
-Block 4 finds temporal parameter intervals of decrease of signal intensity.
-Block 5 locates absolute maxima of $S_{(t_j)}$ on temporal intervals of decrease of signal intensity.
-Block 6 finally determines boundaries of open syllable and boundaries between sounds in a syllable using data obtained from blocks 3 and 5.
Thus in the output of the system we find boundaries of open syllables and sound structures in every syllable. Accuracy of boundary measurements is determined by discrete time quantization of data transfered from separater of informative parameters (block 1).
As it has been demonstrated by numerous experiments, phonotactic information is of great importance. In a number of automatic speech recognition systems such information was not taken into account, which significantly reduced the percentage of correctly recognized words in continuous speech. Information of phoneme concatenation in speech chain forms a filter which passes actual combinations of phonemes and blocks phonotactically impossible ones. Information of phoneme concatenation inter morpheme and word junctures is also of great significance.
The characteristic pronunciation features, phonologic, accentual and rhythmic peculiarities of the utterance in different languages can lead to certain constraints and additional complications in detecting syllabic boundaries. The difficulties increase in case of languages with a high amount of consonants in speech continuum, in which the correct extraction of syllables helps to solve the task of correct recognition and understanding of a spoken message.
Thus, three-level segmentation of continuous speech is proposed: first, (CCC...) CV-syllables are marked along with the exact definition of their boundaries, then syllables are segmented into certain sound types. At the final stage the boundaries of linguistic units are specified on the basis of phonotactics.
Usage of additional acoustic data about the way and place of sound formation allows to define sound structure of segmented syllables based on spectral, temporal and energy parameters which are rather easily and reliably separated by special devices or algorithmic.

REFERENCES

1. T.A.Barašova, B.N.Rudnyi, V.N.Trunin-Donskoj,"Ob avtomatičeskoj segmentacii rečevogo potoka pri vvode rečevogo signala s ustrojstva vydelenija priznakov",V kn.: Rečevoje upravlenije,Moskva,1972.

2. H.Fujisaki, K.Hirose, T.Inone, "Automatic recognition of spoken words from a large vocabulary using syllable templates", IEEE, 1984.

3. C.Gagnoulet, G.Mercier, R.Vives,J.Vaissiere, "A multi-purpose speech understanding system",IEEE,International Conference on Acoustics Speech & Signal Processing,1977.

4. R.K.Potapova, "Avtomatičeskaja segmentacija reči na psevdoslogovyje edinicy" ("Automatic segmentation of speech into pseudo-syllabic units", Proceedings of the first Jntern. Workshop on Natural Communication with Computers, Warsaw,1980.

5. "Urovni jazyka v rečevoj dejatelnosti. K probleme lingvističeskogo obespečenija avtomatičeskogo raspoznavanija reči",Leningrad,1986.

TOWARDS AN AUTOMATIC LABELLING SYSTEM

Charles Barrera
Jacques-François Malet
Nadine Vigouroux

Laboratoire C.E.R.F.I.A.
UA 824-CNRS GRECO
118, Route de Narbonne
31062-TOULOUSE - F R A N C E

Jean Caelen
Geneviève Caelen-Haumont

ICP/INPG
LA 368-CNRS GRECO
46, Avenue Félix Viallet
38031-GRENOBLE - F R A N C E

ABSTRACT

We describe the environment required for a fine frequential labelling; i.e., the code and the operation systems resorted to. We also show how it is possible to devise a system that is capable of assisting an automatic labelling system.

I. INTRODUCTION

Certain problems, connected to Acoustic-Phonetic Decoding, call both for the elaboration of Acoustic Phonetic Data Bases (APDB) and —if only to constitute reference systems— for their labelling. Within the scope of the "Spoken Communication" G.R.E.C.O. (CNRS Coordinated Research Group), various mutually complementing approaches to labelling have been retained; e.g., broad, fine (both temporal and frequential) [1], [2], [3], normative phonetic transcription, etc.. In the present article, we describe the "environment" required for a fine frequential labelling; i.e., the code and operation systems resorted to. We also show how, thanks both to manual labelling and to an APDB, it is possible to devise a system that is capable of assisting an automatic fine frequential labelling. In order to do this, we use phonetic units to set up a correspondence between strips of both signal and spectrum, so that information items —that are useful to both learning and assessment procedures— can in due time be extracted.

II. THE ACOUSTIC MODULE

The first step, in fine frequential labelling, consists in achieving a spectral analysis of the vocal signal. The module of acoustic processing, we have on hand, is derived from a filter bank [4] : it yields spectrum in decibels on a 24-channel MEL scale. One spectral sample corresponds to a 128-dot window of analyzed signal; therefore, over a duration of 8 ms. at a 16 kHz sampling frequency. In practice, this is the one analysis we use, although we have on hand other methods —e.g., FFT, Cepstrum, LPC.

In order both to interpret and identify the signal, the expert makes use of several types of parameters:

-instantaneous values of: i)signal energy in dB (measured immediately after pre-stressing due to the ear-model), ii)formants, iii)spectral cues [5] and iv)fundamental frequency.

-temporal evolution both of the above parameters and of the Continuous/Discontinuous cue that measures the spectral derivative.

Once the parametrization system is specified, there remains to define the unit which the expert is going to work with. The unit we retained is the homogeneous infra-phonemic segment. At this level, processing is entirely automated.

The boundaries of the homogeneous-segment unit are determined through a segmenting function, computed on the basis of the overall variation of the acoustic and prosodic cues —using a modified version of delta coding [6]. A boundary is automatically set, whenever the segmenting function happens to exceed a certain threshold value —that is variable and decreasing with time— or whenever the unit exceeds 60 ms. in duration. The unit, thus obtained, is bound to remain smaller than the phoneme, whatever the value reached by the segmenting function.

III. LABELLING METHODOLOGY

Labelling consists in placing a set of codes, either directly onto the signal in the case of temporal labelling, or onto the spectrogram in the case of frequential labelling. True enough, the temporal domain is still favored by phoneticians: the raw signal is devoid of mathematical processing, that is an unfailing source of alterations. Nevertheless, we chose spectral reading:

-on the one hand, within a spectrum, phenomena such as nasal murmur or friction can both be detected,

-on the other hand, the automated definition of homogeneous segments renders the phonetician's task easier.

The codes to be placed can concern a whole segment [2] or, again, they can be used to spot events accurately [1]. Our approach is segmentwide: the expert places a set of labels —whose definition is underpinned by a phonetic model— at the boundaries of the "automatic" segments, defined above.

III.1 The System of Codes

III.1.1 Definition of Label Vector-Components

Acoustic, phonetic and syllabic properties are characterized through a label vector that is made up of several components, placed according to a previously set positional order. The system of codes consists of six different classes; five of which are set and only one allowed to vary:

-two classes —macro-class, C1, and phoneme code, C2— aim at phoneme characterization,

-two other classes, more closely related to the homogeneous segment, help in further specifying C1 and C2 above. The class dealing with acoustic phonetic modality, C6, is left to vary (i.e., several simultaneous descriptive adjectives are allowed). Class C4 consists of contextual attributes, all coined to give an account of co-articulation phenomena.

-a further class —acoustic phases, C3—sets segments in sequence within a given phoneme realization,

-a final class, C5, supplies information at the syllable level, depending upon the position occupied by a syllable within larger conceptual entities; e.g., word, wordgroup, phrase.

Let us now take up each such class, in the order the expert follows while labelling:

C1: Macro-Class

We recognize ten distinct macro-classes: Vowel (V), Nasal Vowel (M), Semi-vowel (W), Liquid (L) covering /l/ and /r/, Nasal Consonant (N), Voiceless Occlusive (Q), Voiced Occlusive (O), Voiceless Fricative (S), Voiced Fricative (Z) and Silence (P).

C2: Phoneme Codes

These codes take after the International Phonetic Alphabet (IPA), which are not available on computer keyboards; e.g., /ã/ coded AN, /ɔ/ coded O, etc.

Class C1 is redundant, with respect to the phoneme code C2; indeed, it is automatically generated by the system, whenever the expert identifies a phoneme.

C3: Acoustic Phase

This component describes the temporal unfolding of homogeneous segments within a given phoneme: Onset or Establishment phase (E), Sustained or Steady phase (T) and Coda or Phaseout (Q). These three phases are applied systematically to both consonants and vowels.

C4: Contextual Attributes

These describe how contextual events concur with or, sometimes even, prevail over the expected realization of a phoneme; staking phenomena that often more properly pertain to phonology. The cases most often encountered are: A for Approximating (e.g., final /ə/), R for Substituting, I for Insert, F for Merging.

C5: Syllable Delimitors

These codes are defined over non-complex phrases that are considered in a twofold manner:

-along the axis of structural complexity: Phrase (P), Groupword (G), Syllable (I, S, F, M),

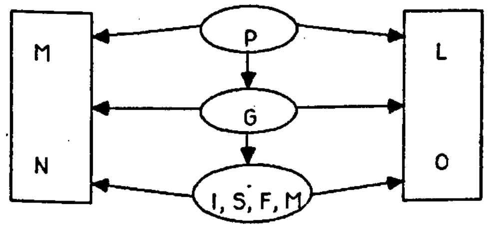
-along the lexical axis: L for Lexical word, O for Tool or Grammatical word.

An additional distinction is made between mono- (M, N) and pluri-syllabic (L, O) words. Examples:

PL: first syllable within a phrase that begins with a lexical word consisting of more than just one syllable,

GM: first syllable within a wordgroup that begins with a lexical word consisting of more than just one syllable,

IL: first syllable within a lexical word.



C6: Modality of Realization

Acoustic or articulatory modality specifies some of the implicit features pertaining to a macro-class (loss and/or addition and/or alteration of acoustic features). Since the field of modality is open to variability, it is possible to choose among a number of descriptive adjectives: Oral (O), Vocalic (V), Glottal (G), Nasal (N), Consonantal (C), Unvoiced (S), Semi- (2), Closure (K), Burst (X), Fricative (F), Palatalized (Y), Affricate (Z), Aspirate (H), Noisy (B). This list is by no means exhaustive, and can be updated should new acoustic features become pertinent.

Defining and placing these label-components appeared, at first, rather to be a matter of interpretation than one of description. However, little by little, there began to emerge a number of steady conventions, likely to sustain a more constant phonetic interpretation of acoustic facts; thus resulting in a type of labelling that is more descriptive than interpretative. Still, classes C1, C2 and C3 can be considered as belonging rather to the descriptive type, whereas C4 and C6 are definitely more subjective and are, therefore, a matter of interpretation. More specifically, Acoustic Phase (C3) is:

-at times, descriptive, and this is the case both of discontinuous vowels and of discontinuous consonants, both fairly easily opened to segmenting rules,

-at other times interpretative, and this is the case —delicate, if anything— of semi-vowels, over which the notion of phase applies with truly extreme difficulties.

III.1.2 Verifying the Labels

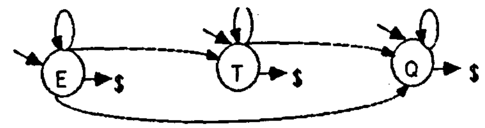
Errors, in implementing both the syntax and the semantics of labelling, can intervene in the course of manual labelling. Therefore, it is necessary to check the manually applied labels, at least for proper syntax.

The procedure is run in three steps:

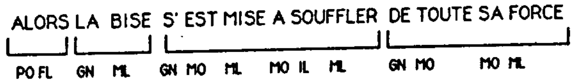
-Since the label vector is a pre-defined structure of components both belonging to a finite set of values and obeying to a strict positional order, it is procedurally possible to check that the value, specifically assigned to a component, does belong to the appropriate set of definition of such values. Thus typing mistakes, which necessarily cause improper labels to be entered, can be automatically detected and then removed.

-Within the temporal sequence, while shifting from one label to the next, choice of value is not arbitrary; indeed, it is subjected to sets of rules respectively applying to the different types of components. Successive values are not drawn, from one such set, in a random order. Thus, for example, when labelling for acoustic phase (although, in practice, labelling order is accross all class-defining sets to formulate one label at a time), sequences such as [E Q T] or [E T Q E], ..., over one and the same phoneme, are strictly prohibited.

The process rule can be symbolized through the following automaton:



Likewise, sequencing syllable codes within phrase structure can be managed by a similar automaton. Examples of syllable coding from the corpus of the G.R.E.C.O.'s Database of sounds [7]:



-The third verification step bears upon mutual compatibility among the components making up one and the same label and, more particularly, on combinations involving C1-C2-C3. Unauthorized combinations are those involving either a redundant modality (Ex.: VO; i.e., vowel V-Oral modality) or a contradictory one (Ex.: by definition, macro-class "Q" excludes Oral modality "O").

III.2 The Operating System

III.2.1 The Spectrogram Editor

Labels placed by an expert are machine-acquired thanks to a spectrogram editor. This software makes it possible to listen to the signal, to view it, to display the corresponding spectrogram, as well as the various cue curves —viewable with or without zoom, to watch both formant values and fundamental frequency, ...

Such a system offers two major advantages:

a)Default labelling is necessarily infra-phonemic, since the labelling agent must set his/her/its view of reality in correspondence with the automatically determined segments; that is, in the case of fine labelling. The system is capable, as well, restrictively to deal with only phoneme designation; labelling over classes C1 and C2 only. Thus it is possible to label broadly (Likewise, it would be possible to concentrate on supra-phonemic units; e.g., diphones).

b)The editor can handle any system of codes. Two other systems are, at present, being used within the scope of specific studies (foreign languages [8] and phrase complexity [9]). A user has only to specify the structure of the wanted label vector, the set of codes to be used by each type of component and, of course, their syntax within a given datastructure.

III.2.2 Verification Procedures

Once labelling is over, a verification procedure is initiated on labels. Procedures, defined on the basis of concepts mentioned earlier (See III.1.1 supra), supply an opportunity for a correction that is interactive with the user. Such a module ensures both a quality- and a reliability-control of the labels produced for the database.

III.2.3 Recapitulating Example

In order to label, the expert has on hand:

-signal that can be both viewed and listened to,

-information items, displayed as spectrograms and curves. Description of these items: (cf. Figure 1)

From left to right, we can see:

-N, the spectrum sample number,

-W, the signal's energy in dB, with an evolution curve vs. time,

-homogeneous segments —whose boundaries are materialized by a number of "<"— are secured through automated segmentation.

-the 8 ms. skeleton spectrum,

-the cues, mentioned supra, displayed as histograms,

-the segmenting marks, properly speaking, that the expert places in correspondence with homogeneous segment (A very high proportion of automatically delivered homogeneous segments get one such mark).

Various zoom-pictures (cues, signal,...) are available, by request, on graphic screen.

IV. THE ACOUSTIC-PHONETIC DATABASE AND LABELLING

We now look at how labelling is closely knit in the elaboration of an APDB.

IV.1 Setting up an APDB

IV.1.1 Information Retained

In order to meet the goals, entailed in setting up a system that delivers automated fine frequential labelling, the necessary APDB must, in the course of manual labelling (system priming), assemble all the required information; namely, information encompassing all processing phases, from the physics of vocal signal to sophisticated linguistic notions.

Two kinds of information, however, should be distinguished:

-Quantitative data: signal samples, spectral samples, prosodic parameters (cf. § II above) and infra-phonemic segment boundaries.

-Qualitative data: labels corresponding to linguistic conceptual events that the expert detects in the course of manual labelling.

All such information is woven into the APDB, thanks to a management system [10].

IV.1.2 Relations

The management system aims at tying together the various types of information, described above.

Through a new datastructure, various kinds of data become associated. For example: signal block number, spectrum sample number, label vectors as placed by expert... Thus, thanks to semantic links, items of symbolic information (phonetic concepts) become associated with items of acoustic quantitative information (spectrum, signal). In fact, this linking correspondence is one of the most crucial problems facing phonetic decoding. At least, this is an important assumption that the database scheme we propose, attempts to meet. Moreover, since users need to retrieve phonetic concepts from any context, it becomes useful to weave relations between the context that is being examined and other previous or subsequent contexts.

Thus, for example, given a phoneme it is possible to find out:

-its realized occurrence among blocks in the signal file,

-its occurrence within the centi-seconds of the spectrum file,

-its phonemic context, both prior and posterior,

-its next realization within the same file.

The latter two types of relations are systematically created for each label vector component. This retrieval scheme, and the set of links it entails, lies at the very base of any APDB consultation.

IV.2 Consulting the APDB

In order to learn an automated labelling system, retrieval of contents from various types of file is imperative (e.g., physical sounds, spectrum, labels issued from expertise, ...). Therefore, to the effect of facilitating new applications —i.e., learning procedures— we designed a retrieval sytem to reach all information elaborated from the vocal signal. This type of consultation is made possible, thanks to the semantic links that allow access both to units pre-defined through labelling and to relations between various datatypes. Thus, it becomes realistic to set up references such as:

-mean value of energy parameter over realizations of phoneme /i/ within a given corpus, uttered by a given speaker,

-mean value of fricative formant over realizations of F-class vowels (displaying the fricative modality),

-mean value of energy parameter over all **T** phases (sustained portion) of all occlusives within corpus,

-etc.

IV.3 Database Contents

At present, we have available an initial acoustic-phonetic database, labelled for 10 speakers (7 males, 3 females). The corpuses used are:

-connected digits and logatoms CVCVCV, both for C.N.E.T. Agreement [11],

-continuous speech: "La Bise et le soleil", for G.R.E.C.O., for a total of 13000 phonemes, labelled.

V. TOWARDS AN AUTOMATED LABELLING

The goal we set for ourselves is to help the phonetician's expert work. We mean either to automatize certain tasks or to further the degree of automatization, already achieved within the pre-segmentation module that yields homogeneous segments.

As an initial step, we limit our scope to the identification of both phoneme (C2) and modality (C6) components of a label vector. In the way of system input, we already have a normative phonetic transcription and a set of quantitative items of information (spectrum, signal) concerning any sequence we wish to label. From this transcription, we contemplate both introducing automated alignment procedures [12], [13], [14], [15] and comparing these with procedures that segment for events [16].

By automatically placing boundaries, such procedures should make it possible to delimit phonemes. Meanwhile, for the purpose of fine labelling, it is equally advantageous to add procedures for extracting acoustic and phonetic features (C6). Specifications for this phase must include:

-not only a strategy of expert labelling [17], [18], [19],

-but also learning results delivered by statistical modules, when these are run on a base of already labelled data.

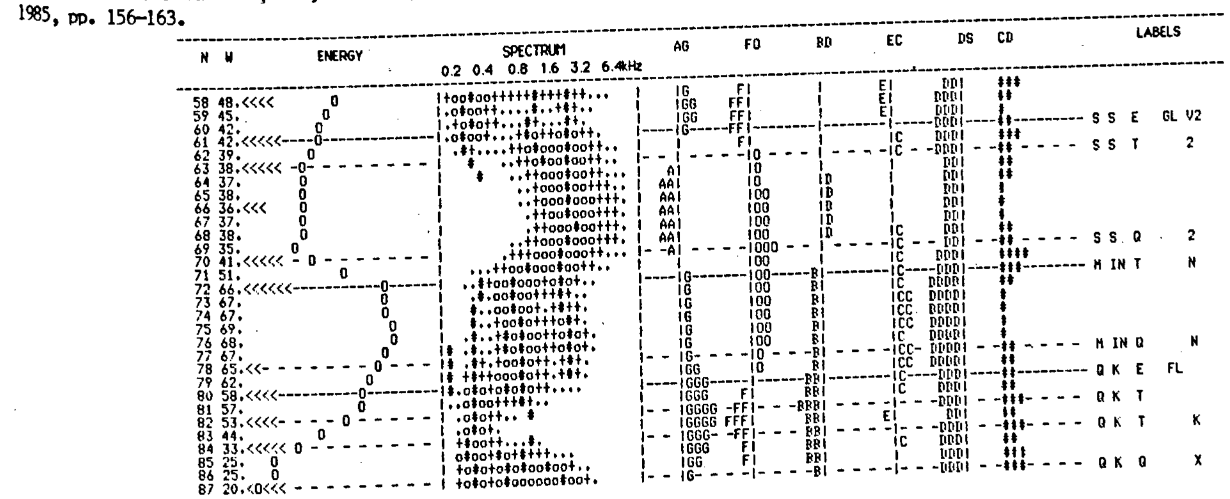
For the time being, the system is devised both to validate labels, and as a tool serving expertise. Next, we mean to formalize our results, with a view to elaborating an automated interactive labelling sytem.

ACKNOWLEDGMENT
Our warm thanks to Dany Laur who joined us on the lengthy spectrogram-reading expert's task.

VI. BIBLIOGRAPHIC REFERENCES

[1] C. Abry, C. Benoît, L.J. Boé, R. Sock, "Un Choix d'Evénements pour l'Organisation Temporelle du Signal de Parole", **XIV JEP, GALF-CNRS**, Paris, Juin 1985, pp. 133-137.
[2] D. Autesserre, M. Rossi, "Propositions pour une Segmentation et un Etiquetage Hiérarchisé. Application à la Base de Données Acoustiques du GRECO- CP", **XIV JEP, GALF-CNRS**, Paris, Juin 1985, pp. 147-151.
[3] C. Abry, D. Autesserre, C. Barrera, C. Benoît, L.J. Boé, J. Caelen, G. Caelen-Haumont, M. Rossi, R. Sock, N. Vigouroux, "Propositions pour la Segmentation et l'Etiquetage d'une Base de Données des Sons du Français", **XIV JEP, GALF-CNRS**, Paris, Juin 1985, pp. 156-163.
[4] J. Caelen, "Space/Time Data-Information in ARIAL Project Ear Model", **Speech Communication**, Vol 4, Aug. 1985, pp. 163-179.
[5] J. Caelen, G. Caelen-Haumont , "Indices et Propriétés dans le Projet ARIAL II", **Actes du Séminaire Encodage et Décodage Phonétiques, GALF-CNRS**, Toulouse, 1981, pp. 129-143.
[6] N. Vigouroux, J. Caelen, "Segmentation Phonétique et Organisation d'une Base de Données Acoustiques et Phonétiques", **XIV JEP, GALF-CNRS**, Paris, Juin 1985, pp. 152-155.
[7] R. Descout, J.F. Sérignat, O. Cervantes, R. Carré, "Une Base de Données des Sons du Français", **12-th ICA**, July1986.
[8] J.F. Malet, "Une Méthode Acoustico-Phonétique pour l'Enseignement Automatique de Langues Etrangères", **XV JEP, GALF-CNRS**, Aix-en Provence, Mai 1986.
[9] G. Caelen-Haumont, "Grammatical Components and Macro-Prosody: Quantitative Analysis Toward Statistical Correlations", **12-th ICA** , Toronto, July 1986.
[10] J. Caelen, N. Vigouroux, "An acquisition and Research System for an Evolving Nucleus of Acoustico-Phonetic Knowledge", **IAPR, 8-th ICPA, AFCET**, Paris, 28-31 Octobre 1986.
[11] **Convention CNET N° 86 7B 020**, "Réalisation et Exploitation d'un logiciel de validation d'indices Acoustiques pour la Reconnaissance de la Parole Multi-Locuteur".
[12] J.S. Bridle, R.M. Chamberlain, "Automatic Labelling of Speech Using Synthesis-By-Rule and Non-Linear Time-Alignment", **11-th ICA**, Toulouse 1983, pp. 187-189.
[13] A. Andreewsky, M. Desi, C. Flur, F. Poirier, "Une méthode de Mise en Correspondance d'une Chaîne Phonétique et de sa Forme Acoustique", **11-th ICA**, Toulouse 1983.
[14] P. Collins, S. Barber, "Fine Phonetic Labelling Methodology for Speech Recognition Research", **Proceedings IEEE-ICASSP**, Tokyo 1986, pp. 2779-2782.
[15] H.C. Leung, V.W. Zue, "A procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech", **Proceedings IEEE-ICASSP**, San Diego, 1984, pp. 2-9.
[16] G. Pérennou, M. de Calmes, "Segmentation en événements phonétiques et unités syllabiques", **XIV JEP, GALF-CNRS**, Paris, Juin 1985, pp. 142-146.
[17] F. Lonchamp, "Reading Spectrograms: The Viewfrom the Expert", in **Fundamentals in Computer Understanding: Speech and Vision**, ed. J.P. Haton, Cambridge University Press, 1987, pp.181-206.
[18] N. Carbonnel, D. Fohr, J.P. Haton, F. Lonchamp, J.M.Pierrel, "An Expert-System for the Automatic Reading of French Spectrograms", **Proceedings IEEE-ICASSP**, San Diego 1984, pp. 42-8.
[19] P.E. Stern, M. Eskenazi, D. Memmi, "An Expert System for Speech Spectrogram Reading", **Proceedings IEEE.-ICASSP**, Tokyo,1986.

FIGURE 1



VOWEL RECOGNITION BASED ON "LINE-FORMANTS" DERIVED FROM AN AUDITORY-BASED SPECTRAL REPRESENTATION*

Stephanie Seneff

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

A new approach to vowel recognition is described, which begins by reducing a spectrographic representation to a set of straight-line segments that collectively sketch out the formant trajectories. These "line-formants" are used for recognition by scoring their match to a set of histograms of line-formant frequency distributions determined from training data for the 16 vowel categories in the recognition set. Speaker normalization is done by subtracting F_0 from line-formant frequencies on a Bark scale. Although the formants are never enumerated or tracked explicitly, the frequency distributions of the formants are the main features influencing the recognition score. Recognition results are given for 2135 vowels extracted from continuous speech spoken by 292 male and female speakers.

INTRODUCTION

The formant frequencies are probably the most important information leading to the recognition of vowels, as well as other sonorant and even possibly obstruent sounds. Therefore, researchers have spent a considerable amount of effort designing robust formant trackers, which attempt to associate peaks in the spectrum with formant frequencies, using continuity constraints to aid in the tracking of the formants. Once the formant tracks are available, it then becomes possible to identify directions and degree of formant movements, features that are important in recognizing diphthongs, semivowels, and place of articulation of adjacent consonants.

It is impossible to design a "perfect" formant tracker. The most serious problem with formants is that when they are wrong there are often gross errors. Therefore, we have decided to adopt a somewhat different approach, one that can lead to information about formant movements without explicitly labelling the formant numbers. The method also collapses the two stages of formant tracking and track interpretation (e.g., "rising formant") into a single step. The outcome is that a spectrographic representation is reduced to a skeleton sketch consisting of a *set* of straight-line segments, which we call "line-formants," that collectively trace out the formant tracks. The recognition strategy then involves matching all of the line-formants of an unknown segment to a set of templates, each of which describes statistically the appropriate line-formant configurations for a given phonetic class (which could be as detailed as "nasalized /æ/" or as general as "front vowel"). Usually the number of line-formants for a given speech segment is considerably larger than the number of formants, because in many cases several straight-line segments are required to adequately reflect the transitions of a single formant.

*This research was supported by DARPA under Contract N00039-85-C-0254, monitored through Naval Electronic Systems Command.

SIGNAL PROCESSING

Spectral Representation

The system makes use of two spectrogram-like representations that are based on our current understanding of the human auditory system. These have been described in detail previously [1,2], and will only be discussed briefly here. The analysis system consists of a set of 40 critical band filters, spanning the frequency range from 160 to 6400 Hz. The filter outputs are processed through a nonlinearity stage that introduces such effects as onset enhancement, saturation and forward masking. The outputs of this stage are then processed through two independent analyses, each of which produces a spectrogram-like output. The "Mean Rate Spectrogram" is related to mean rate response in the auditory system, and is used for locating sonorant regions in the speech signal. The "Synchrony Spectrogram" takes advantage of the phase-locking property of auditory nerve fibers. It produces spectra that tend to be amplitude-normalized, with prominent peaks at the formant frequencies. The amplitude of each spectral peak is related to the amount of energy at that frequency *relative* to the energy in the spectral vicinity. The line-formant representation is derived from this Synchrony Spectrogram.

Line-formant Processing

The line-formants are obtained by first locating sonorant regions, based on the amount of low frequency energy in the Mean Rate Spectrogram. Within these sonorant regions, a subset of robust peaks in the Synchrony Spectrogram is selected. Peaks are rejected if their amplitude is not sufficiently greater than the average amplitude in the surrounding time-frequency field. For each selected peak, a short fixed-length line segment is determined, whose direction gives the best orientation for a proposed formant track passing through that peak, using a procedure as outlined in Figure 1. The amplitude at each point on a rectangular grid within a circular region surrounding the peak in question is used to update a histogram of amplitude as a function of the angle, θ. Typical sizes for the circle radius are 20 ms in time and 1.2 Bark in frequency. The maximum value in the histogram defines the amplitude and corresponding θ for the proposed track, as marked by an arrow in Figure 1*c*.

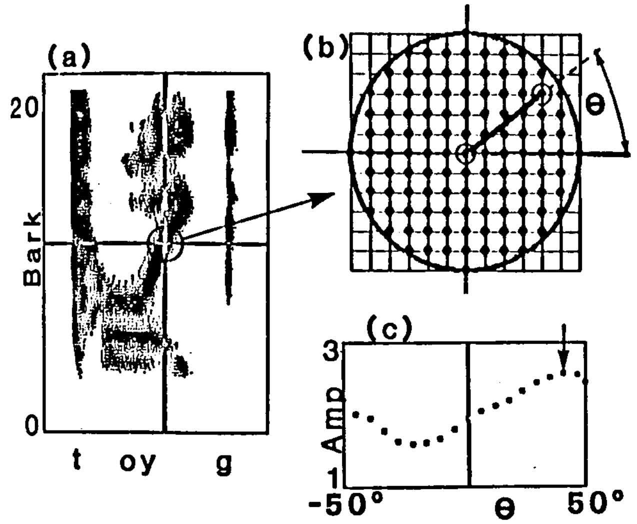


Figure 1: Schematic illustration of process used to determine an orientation for a formant passing through a peak. (a) Synchrony Spectrogram with cross-bars indicating a referenced peak. (b) Schematic blow-up of region around the peak, outlining procedure to generate a histogram of amplitude as a function of angle. (c) Resulting histogram for the example in part *a*.

At each time frame several new short segments are generated, one for each robust spectral peak. A short segment is then merged with a pre-existing partial line-formant whenever the two lines have a similar orientation, and the distance between each endpoint and the other line is sufficiently small. The merging process is accomplished by creating a weighted-average line-formant that incorporates the new line. If a given new segment is sufficiently unique, it is entered as a new partial line-formant.

The resulting *Skeleton Spectrogram* for the /ɑ/ in the word "shock" is illustrated in Figure 2*a*, along with a *Schematized Spectrogram* in Figure 2*b*, included to facilitate visual evaluation. The latter is constructed by replacing each line-formant with a time sequence of Gaussian-shaped spectral peaks with amplitude equal to the line's amplitude. The corresponding Synchrony Spectrogram is shown in Figure 2*c*, with line-formants superimposed. For direct comparison, Figure 2*d* shows a Synchrony Spectral cross section at the time of the vertical bar, on which is superimposed a cross section of the Schematized Spectrogram. For this example, we see that peak locations and amlitudes in the vowel are accurately reflected. In addition, formant transitions appropriate for the palatal fricative on the left and the velar stop on the right are also captured.

RECOGNITION EXPERIMENT

Thus far, we have focused our studies on speaker-independent recognition for 16 vowels and diphthongs of American English in continuous speech, restricted to obstruent and nasal context. The semivowel context is excluded because we believe that in many cases vowel-semivowel sequences should be treated as a single phonetic unit much like a diphthong.

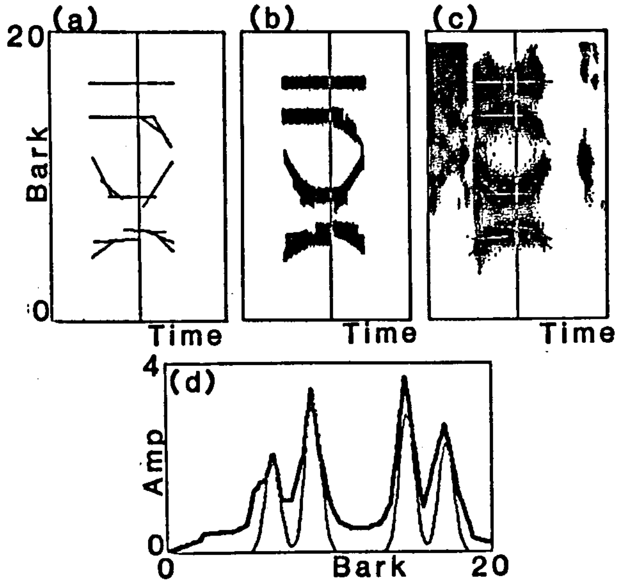


Figure 2: Sample line-formant outputs: (a) Skeleton Spectrogram for word "shock," (b) Corresponding Schematized Spectrogram, (c) Synchrony Spectrogram with line-formants superimposed, (d) cross-sections from *b* and *c* at the cursor, superimposed.

Speaker Normalization

Our first task was to devise an effective speaker-normalization procedure. Many investigators have noted the strong correlation between formant frequencies and F_0 [3]. The relationship is clearly nonlinear – the second formant for female /i/ is higher on average by several hundred Hz, whereas the F_0 difference is on the order of 100 Hz. However, on a Bark (critical band) scale the male-female difference in F_2 for /i/ becomes much more similar to that in F_0. Thus we decided to try a very simple scheme – for each line-formant, subtract from the line's center frequency the *median* F_0 over the duration of the line, on a Bark scale.

We found this normalization procedure to be remarkably effective, as illustrated in Figure 3. Part *a* shows a histogram of the center frequencies of all of the lines for 35 male and 35 female /æ/ tokens. Part *b* shows the same data, after median F_0 has been subtracted from each line's center frequency. The higher formants emerge as separate entities after the F_0 normalization. The normalization is not as effective for F_1, but the dispersal in F_1 is due in part to other factors such as vowel nasalization. A valid question to ask is the following: if it is supposed that speaker normalization can be accomplished by subtracting *a factor times* F_0 from all formant frequencies, then what should be the numerical value of the factor? An answer can be obtained experimentally using autoregressive analysis. We defined $F_n' = F_n - \alpha F_0$ to be the normalized formant frequency for each line. Using vowels for which the formants are well separated, we associated a group of lines with a particular formant such as F_2. The goal was to minimize total squared error for each remapped formant among all speakers, with respect to α. The resulting estimated value for α was 0.975, providing experimental evidence for the validity of the proposed scheme.

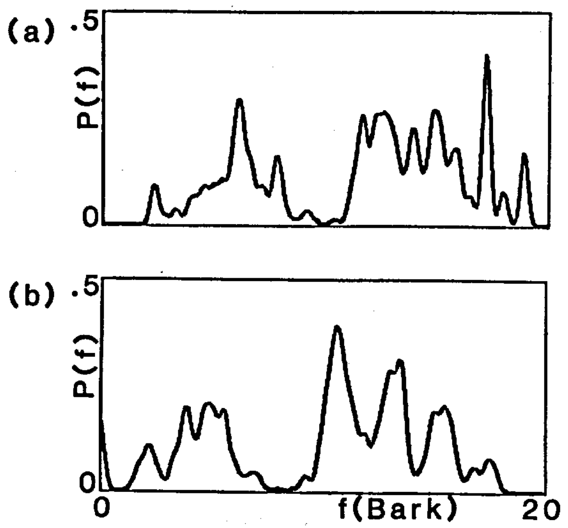


Figure 3: Histograms for center frequencies of all line-formants for 35 female and 35 male tokens of /æ/, (a) without F_0 normalization, and (b) with F_0 normalization.

Scoring Procedures

Our goal in developing a recognizer for the vowels was to emphasize the formant frequency information without ever explicitly identifying the formant numbers. We wanted to avoid traditional spectral template-matching schemes, because they depend too heavily on irrelevant factors such as the loudness or the overall spectral tilt. On the other hand, we did not want to specify, for example, the distance between F_2 and a target F_2, because this relies on accurately enumerating the formants.

We decided to construct histograms of frequency distributions of spectral peaks across time, based on data derived from the line-formants. The scoring amounts to treating each histogram as a probability distribution, and matching the unknown token's line-formants against the appropriate distributions for each vowel. To construct the histograms for a given vowel, all of the line-formants in a training set were used to generate *five* histograms intended to capture the distributions of the formants at significant time points in the vowel. All lines were normalized with respect to F_0, which was computed automatically using a version of the Gold-Rabiner pitch detector [4]. Each line-formant's contributions to the histograms were weighted by its amplitude and its length.

Only left, center and right frequencies of the lines were used in the histograms. The left frequency of a given line-formant falls into one of two bins, depending upon whether or not it is near the beginning of the vowel. Right frequencies are sorted similarly, with a dividing point near the end of the vowel. Center frequencies are collected into the same histogram regardless of their time location. Such a sorting process results in a set of histograms that reflects general formant motions over time. For example, the F_2 peak in the histograms for /e/ shifts upward from left-on-left to center to right-on-right, reflecting the fact that /e/ is diphthongized towards a /y/ off-glide, as illustrated in Figure 4.

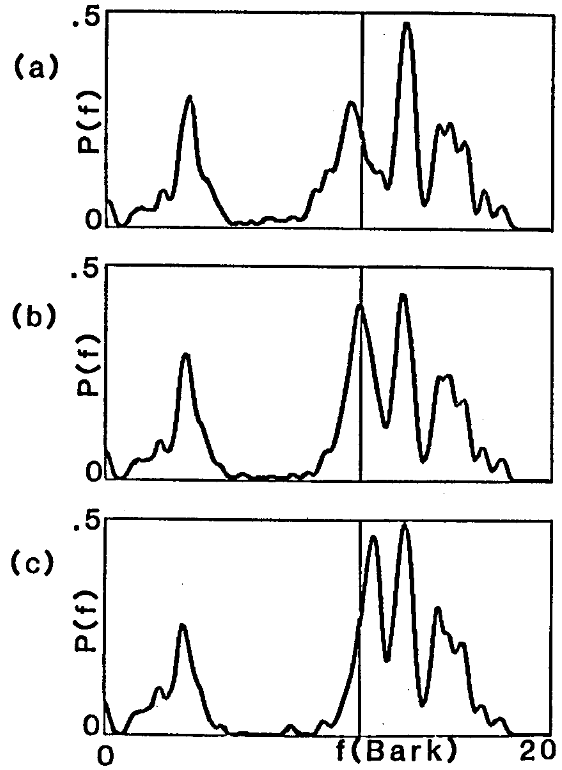


Figure 4: Histograms for (a) left-on-left, (b) center, and (c) right-on-right line-formant frequencies for 128 tokens of /e/, F_0 normalized.

2135 Vowels, 288 Speakers

ü	i	ɪ	e	ɛ	æ	aʷ	aʸ	ɑ	ʌ	ɔ	ɔʸ	o	ʊ	u	ɝ
90	220	268	128	153	155	131	92	103	147	156	96	83	114	96	103

Table 1: Distributions of vowels in recognition experiment

To score an unknown token, the left, center, and right frequencies of all of its lines are matched against the appropriate histograms for each vowel category, which are treated as probability distributions. The score for the token's match is the weighted sum of the log probabilites for the five categories for all of the line-formants. The amplitude of the line does not enter into the match, but is used only as a weight for the line's contribution to the score. This strategy eliminates the problem of mismatch due to factors such as spectral tilt or overall energy.

Recognition Results

The vowels used for recognition were extracted from sentences in the TIMIT database [5]. The speakers represented a wide range of dialectical variations. A total of 2135 vowel tokens spoken by 206 male and 82 female speakers were used as both training and test data, using a jackknifing procedure. The distributions of vowels are shown in Table 1. Each speaker's vowel tokens were scored against histograms computed from all of the line-formants *except* those from that speaker. The scoring procedure was as discussed above, with histograms defined for sixteen vowel categories. The endpoints for the vowels were taken from the time-aligned phonetic transcription.

	ü	i	ɪ	e	ɛ	æ	aʷ	aʸ	ɑ	ʌ	ɔ	ɔʸ	o	ʊ	u	ɝ
ü	49	15	11	6	3	1								4	8	3
i	11	70	5	8	1					1			1	2	1	
ɪ	11	10	32	16	11	5				2				6	5	2
e	5	9	5	60	7	6	1	1		1		3		2		
ɛ	3	1	12	14	37	16	1	1		6			2	3	1	4
æ	1	1	1	10	9	59	4	7		4		1		1	2	1
aʷ					2	13	39	6	13	7	7	2	7	3		2
aʸ					1	7	4	58	15	2	1	8	2	2		
ɑ						3	7	15	40	4	27		5			
ʌ	1	1	1		7	6	2	5	17	39	2	2	5	6	5	1
ɔ							3	3	29	1	46	4	12	1	1	1
ɔʸ						1	1	10	1		3	67	8		6	2
o						4	7	1	5	6	14	4	53		5	1
ʊ	20	1	11		4	1			1	5	3	3	9	28	11	4
u	11	2	2	2	1	1	1	1	2	4	1	2	9	17	40	3
ɝ	8		4	3	3		1	1	1	3	1	1	5	5	3	62

Table 2: First choice confusion matrix for the vowels
Row = Labeled Category, Column = Recognized Category.

A matrix of first-choice confusion probabilities is given in Table 2, in terms of percent correct in the phonetic category. For the most part, confusions are reasonable. We feel encouraged by this performance, especially considering that multiple dialects and multiple contexts are included in the same histogram.

Figure 5 summarizes recognition performance in terms of percentage of time the correct answer is in the top N, for all speakers, and for male and female speakers separately. Recognition was somewhat worse for females, who represented only 25% of the population. Also shown are the recognition results for female speakers when the F_0-normalization scheme is omitted, both in collecting the histograms and in scoring. Significant gains were realized as a consequence of the normalization. The performance for the male speakers without F_0 normalization however (not shown) did not change.

FUTURE PLANS

We believe that recognition performance can be improved by extensions in several directions. One is to divide each vowel's histograms into multiple subcategories, based on both general features of the vowel and coarticulation effects. General categories, useful for the center-frequency histogram, would include "nasalized," "Southern accent," or "fronted." Left- and right-context place of articulation, such as "velar," could be used to define corresponding histogram subcategories. We also plan to explore an alternative recognition strategy for explicitly matching each *line-formant* against a set of template *line-formants* describing a particular phonetic category, instead of reducing the line to three "independent" points. We believe that such an approach will better capture the fact that a given left frequency and a given right frequency are connected. Finally, we plan to gradually expand the scope of the recognizer, first to vowels in all contexts and then to other classes such as semivowels.

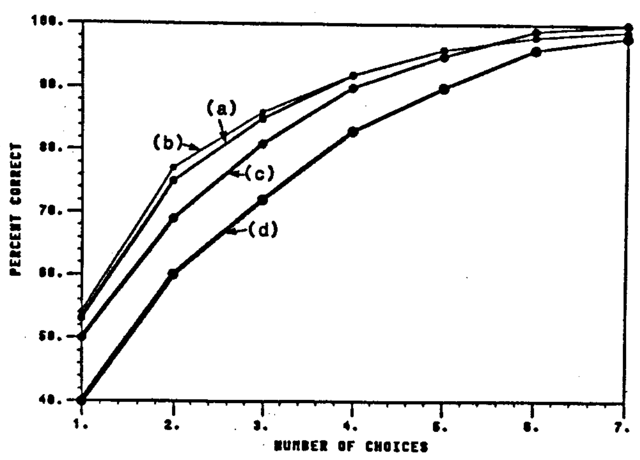


Figure 5: Recognition results expressed as percent of time correct choice is in top N, for the following conditions: (a) all speakers, (b) males only, (c) females only, and (d) females without F_0 normalization.

REFERENCES

[1] Seneff, S. (1986) "A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research," ICASSP Proceedings, Tokyo, Japan, 37.8.1–37.8.4.

[2] Seneff, S. (1988) "A Joint Synchrony/Mean-rate Model of Auditory Speech Processing," Journal of Phonetics, Special Issue on Representation of Speech in the Auditory Periphery, to appear in Jan.

[3] Syrdal, A. K. (1985) "Aspects of a Model of the Auditory Representation of American English Vowels," Speech Communication 4, 121-135.

[4] Gold, B. and L.R. Rabiner (1969) "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Am. 46, 442-448.

[5] Lamel, L. F., R. H. Kassel, and S. Seneff (1986) "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," Proceedings of the DARPA Speech Recognition Workshop Palo Alto, CA., Feb 19-20, 100-109.

RELIABILITÄTSMASSE FÜR DIE AUTOMATISCHE TRANSKRIPTION

REINHOLD GREISBACH

Institut für Phonetik
Universität zu Köln
Greinstr. 2, D-5000 Köln 41

ZUSAMMENFASSUNG

Die Beurteilung der Leistungsfähigkeit von automatischen Transkriptionsverfahren verlangt nach Methoden, die die Validitätsproblematik bei phonetischen Transkriptionen berücksichtigen können. Reliabilitätsmaße scheinen diese Voraussetzung zu erfüllen. Bereits im Hinblick auf zukünftige Entwicklungen wird deshalb hier die Anwendbarkeit solcher Maße auf automatisch erstellte Transkripte theoretisch untersucht.

MOTIVATION

Bereits heute ist es möglich, akustische Sprachsignale automatischen Spracherkennungsprozessen zu unterziehen mit dem Ziel, eine segmentale phonetische (impressionistische) Transkription dieses Sprachsignals zu gewinnen. Dabei läßt jedoch meist ein einziger Blick auf das Resultat erkennen, wie gut oder besser wie schlecht diese automatische Transkription (aT) arbeitet.
Wenn sich jedoch in Zukunft die Leistungsfähigkeit automatischer Erkennungsverfahren weiter verbessert, und daran scheint kein Zweifel, wird die Entscheidung "Transkribiert die Maschine richtig?" nicht mehr durch einfachen Augenschein zu treffen sein.
Bei der automatischen Spracherkennung läßt sich i.a. sehr leicht entscheiden, ob die automatische Erkennung gelingt oder nicht. Dazu wird das Resultat des automatischen Prozesses mit den Höreindrücken einer menschlichen Hörergruppe verglichen. Sehr oft besteht diese Hörergruppe nur aus einer einzigen Person, denn es darf hier vorausgesetzt werden, daß auch jeder andere (muttersprachliche) Hörer das gleiche wahrnehmen würde. Für die Prüfung eines automatisch erstellten Transkripts, das auch mit den Höreindrücken einer menschlichen Hörergruppe verglichen werden muß, ist gerade diese Voraussetzung nicht erfüllt. Denn bei hinreichend enger phonetischer Notation und hinreichend langem Sprachsignal wird unter zwei Transkribenten wohl niemals völlige Einigkeit über das Gehörte herrschen. Dies bedeutet aber, daß die Richtigkeit eines automatischen Transkripts, anders als im Falle der automatischen Spracherkennung, nicht an einem eindeutigen Muster geprüft werden kann. Die Überprüfbarkeit eines automatisch erstellten Transkripts hängt also von der Richtigkeit (Validität) manuell erstellter Vergleichstranskripte ab.
Vor der genaueren Untersuchung der Validitätsproblematik betrachten wir zunächst, wie ein phonetisches Transkript entsteht, und kennzeichnen damit zugleich den Begriff "Transkription", wie er hier verstanden werden soll. Unsere (wohl für die deutschsprachige Phonetik typische) Transkription geht von einem akustischen Sprachsignal aus, welches mithilfe der artikulatorisch definierten IPA-Symbole notiert wird. Der Transkribent verläßt sich dabei ausschließlich auf sein Gehör (auditive Transkription).
Diese Form der Transkription birgt nun eine Besonderheit in sich. Bei der auditiven Transkription mit IPA-Symbolen müssen mögliche artikulatorische Abweichungen des Sprechers kompensiert werden (so erscheint bei der Transkription der Äußerung eines Bauchredners die artikulatorisch nicht vorhandene Lippenrundung eines auditiv wahrgenommenen [y] aufgrund der artikulatorischen Definition dennoch im Transkript.) Notiert werden kann also immer nur ein vorgestellter, ideal artikulierender Sprecher. Umgekehrt bedeutet dies, daß selbst dann, wenn sich die Artikulation des Sprechers (z.B. mittels Röntgenfilm) optisch beobachten läßt, dies noch keine direkte Überprüfung der Notation gestattet.
Was ist also - bei auditiver Transkription mit IPA-Symbolen - eine richtige Notation? Die Antwort ergibt sich durch eine Grenzüberlegung: Wenn für einen Laut jeder Transkribent zu jeder Zeit das gleiche Symbol verwendet, so ist diese Notation richtig (valide)! Die Realität kann jedoch nur aus einer Stichprobe bestehen, die idealerweise verschiedene Transkribenten zu verschiedenen Zeiten von einer Sprachaufnahme durchführen. Und gleichfalls wird diese Stichprobe i.a. nicht für jeden Laut das gleiche Symbol enthalten. Den Grad der Übereinstimmung in einer solchen Stichprobe nennt man Reliabilität, den Grad der Verschiedenheit Variabilität. Reliabilität darf aber grundsätzlich nicht mit Validität verwechselt werden, denn eine maximal reliable auditive IPA-Notation eines von einem menschlichen Sprecher produzierten Sprachsignals ist nur deshalb valide, weil sie auf keine andere Weise als durch das Ohr direkt überprüfbar ist. Bei Notation in anderen Alphabeten, z.B. dem analphabetischen von Jespersen, kann jedoch eine auditive Notation aufgrund von optischen Informationen direkt verifiziert werden (dann nämlich, wenn die artikulatorischen Beschreibungsdimensionen des Alphabets besser den tatsächlichen Bewegungsdimensionen des Artikulationsapparats entsprechen, als dies beim IPA-Alphabet der Fall ist). Für eine eingehendere Diskussion dieser Fragen vgl. /1/. Prinzipiell ist also nur die Reliabilität eines Transkripts feststellbar. Ob damit auch seine Validität bestimmt ist, hängt offenbar vom gewählten Transkriptionsalphabet ab. Für die Bewertung eines automatisch erstellten Transkripts muß die Reliabilität nun zu einer quantifizierbaren Größe werden, zu einem Reliabiliätsmaß.

ÄHNLICHKEITSMASSE

Reliabilitätsmaße oder allgemeiner Reliabilitätsmessungen basieren üblicherweise auf einer numerischen Bewertung der Differenzen zwischen den Symbolen des jeweiligen Transkriptionsalphabets, sog. Ähnlichkeitsmaßen (Ä-Maßen). Anders als den aus dem täglichen Leben vertrauten Maßen der physikalischen Umwelt fehlen den meisten dieser Maße jedoch die (mathematischen) Eigenschaften, die die Bewertung von physikalischen Meßergebnissen einfach gestalten. Welche Eigenschaften ein Maß besitzt, welchem Skalentyp es zuzuordnen ist (wie man in der psychologischen Testtheorie sagt), bestimmt in der Phonetik zuvorderst die Vorstellung, die der jeweilige Phonetiker von den Beziehungen der Symbole untereinander besitzt.
Bei den einfachsten Ä-Maßen sind alle Symbole des Alphabets gleichberechtigt. Sie stehen ohne erkennbare Ordnung nebeneinander, was letztlich bedeutet, daß die Differenz zwischen allen Symbolen gleich groß ist (Nominalskala). Kompliziertere Maße setzen eine Ordnung zwischen den Symbolen voraus, z.B., daß die Differenz zwischen [i] und [e] kleiner ist als die zwischen [i] und [a](Ordinalskala). Bei einer Intervallskala lassen sich darüberhinaus die Differenzen zwischen den Symbolen vergleichen. So ist z.B. bei den Kardinalvokalen der Unterschied zwischen Kardinal-[e] und -[ɛ] definitionsgemäß genauso groß wie der zwischen Kardinal-[ɛ] und - [a].
Um zu einem Zahlenwert zu gelangen, werden zunächst auf der Basis dieser Skalen (ab dem Ordinalskalenniveau) mehrdimensionale Räume konstruiert und die Symbole darin angeordnet. Bei der Konstruktion dieser Räume lassen sich zwei Hauptverfahren feststellen. Der eine Verfahrenstyp geht von den artikulatorischen Klassifikationsdimensionen des Alphabets aus und spannt den Raum entlang dieser Dimension zumeist orthogonal auf (nicht orthogonal z.B. als "Vokaldreieck"). Der andere, aufwendigere Verfahrenstyp erzeugt den Raum mittels auditiver Dimensionen, die nach Hörtests mit Versuchspersonen durch statistische Methoden wie z.B. MDS oder Faktorenanalyse gewonnen werden. Das Ä-Maß gibt dann den Abstand zweier Symbole in diesen Räumen an (mit einer meist heuristisch gewonnenen Abstandsfunktion).
Welcher der Skalentypen ist nun aber für die Transkription der richtige? Die Literatur dokumentiert hier verschiedene Meinungen, wobei die Befürworter des "transkriptorischen Messens" i.a. auf dem Intervallskalenniveau stehen, während sich seine Gegner (konsequenterweise) auf das Nominalskalenniveau zurückziehen (müssen). Alle bekannten Reliabilitätsuntersuchungen auf IPA-Basis (/2/,/3/,/4/,/5/,/6/,/7/,/8/) benutzen Ä-Maße zumindest auf Intervallskalenniveau.

RELIABILITÄTSMASSE

Während die erste quantitative Reliabilitätsuntersuchung zur Transkription bereits zu Anfang dieses Jahrhunderts stattfand /9/, verwenden erst die Arbeiten der '80er Jahre den Begriff "Messen der Reliabilität" bzw. "Reliabilitätsmaß" (R-Maß), der hier (wie in der psychologischen Testtheorie) als Maß für den Grad der Übereinstimmmung je zweier Beobachter (Transkribenten) verstanden wird. Dieser "Korrelationskoeffizient" zweier Transkribenten ergibt sich als die gewichtete Summe aller Unterschiede zwischen den beiden (Stichproben-)Transkripten, gemessen mit dem jeweiligen symbolbezogenen Ä-Maß. Mit einer solchen Messung soll die Befähigung eines Transkribenten für eine transkriptorische Aufgabe festgestellt werden. Er gilt dann als befähigt, wenn seine "Reliabilitäts-Korrelation" zu einer größeren Gruppe /8/ oder zu einem "master-transcriber" /7/ einen bestimmten Grenzwert übersteigt.
Der Transkribent wird damit also zu einem Meßinstrument, dessen Reliabilität (Zuverlässigkeit) meßbar ist. Dieser "Entmenschlichung" des Wissenschaftlers mag es wohl hauptsächlich zuzuschreiben sein, daß Kritik an solchen Reliabilitätsuntersuchungen laut wird /10/.
Es scheint deshalb angeraten, den Begriff Reliabilität anders zu fassen, ihn nicht auf die messende Instanz, sondern auf das gemessene Resultat zu beziehen. So kann man bei einem hohen R-Maß zweier Transkribenten davon ausgehen, daß auch ihre Transkripte zuverlässig, also "reliabel" sind, und ihnen dieses Maß zuweisen (Text-Reliabilität [TR-Maß]). Umgekehrt darf allerdings bei einer geringen Text-Reliabilität nicht gefolgert werden, daß die Transkripte über den gesamten Text gleichmäßig weit voneinander abweichen. Tatsächlich haben die Re-

liabilitätsmessungen der '60er Jahre (/2/, /3/,/4/) gezeigt, daß z. B. die Reliabilität von hohen Vokalen wesentlich höher ist als die von tiefen. Die Meßgröße für die Reliabilität bei diesen Untersuchungen wurde geometrisch/heuristisch gewonnen und läßt sich als Abweichung von einem Mittelwert interpretieren (Lautklassenreliabilität [LR-Maß]). Gibt das LR-Maß quasi paradigmatisch die Reliabilität für jedes Symbol des Alphabets, so kann natürlich auch syntagmatisch jedem Laut des Textes ein solches Maß zugeordnet werden (Symbol-Realibilität [SR-Maß]).
Durch den Übergang vom TR- zum LR- und schließlich zum SR-Maß steigt der Rechen- und insbesondere der Darstellungsaufwand. Andererseits kommt man so der ja eigentlich angestrebten physikalischen Idealvorstellung immer näher, nämlich für jeden einzelnen Meßwert einen eigenen Reliabilitätswert zu bestimmen.
Messen R-Maße die Abweichungen der Notationen verschiedener Transkribenten für eine gegebene Aussprache, so messen Variabilitätsmaße die Abweichungen verschiedener Aussprachen bei ggfs verschiedenen Sprechern. Dafür lassen sich natürlich die gleichen Maße anwenden, die Werte sind nur anders zu interpretieren. Zur Konstruktion und Anwendung eines solchen Maßes auf Wortbasis (Wort-Variabilität [WV-Maß])vgl./12/.

KONSTRUKTIONS- UND ANWENDUNGSPROBLEME

Diese kurze Übersicht dokumentiert, daß sich sehr leicht eine Vielzahl von Ä- und R-Maßen konstruieren läßt (für auch hier verwendbare Maße aus der psychologischen Testtheorie vgl. /11/). Für alle diese Masse bestehen jedoch gewisse gemeinsame Probleme, so daß man die Brauchbarkeit eines Maßes danach beurteilen kann, wie es diesen Problemen gegenübersteht. Es sollen hier einige der augenfälligsten Probleme genannt und zum Teil mit Anmerkungen versehen werden.
(1) Die meisten Ä-Maße (auf Intervallskalenbasis) sehen keine Vergleiche zwischen Vokalen und Konsonanten vor (vgl. /6/,/8/). Für die Ermittlung der Reliabilität muß jedoch manchmal der Abstand von einem Vokal zu einem Konsonanten bestimmt werden. (2) Bei der Konstruktion von solchen Maßen entsteht die Frage nach dem größtmöglichen Abstand im vokalischen und im konsonantischen Bereich. Müssen sie gleich groß sein oder nicht? Eine Reihe ähnlicher Überlegungen im Zusammenhang mit der Konstruktion von Ä-Maßen eskaliert dann in der Hauptfrage: Mit welcher Gewichtung gehen die einzelnen (Klassifikations-)Dimensionen in die Abstandsfunktion ein? (3) Der Übergang vom Ä-Maß zum R-Maß bringt ein neues Problem. Zwei Transkripte des gleiches Sprachsignals werden sicher sehr oft eine unterschiedliche Anzahl von Symbolen enthalten. Somit muß an einigen Textstellen der Abstand von "einem" Laut zu "keinem" Laut bestimmt werden. Die bekannten R-Maße (/6/,/8/) verwenden in diesem Fall den größtmöglichen Abstand. Ein Symbol wird aber doch wohl dann am ehesten im Text fehlen, wenn der jeweilige Laut im Signal undeutlich erscheint, wie etwa bei Reduktionen (z. B. [ra:tən]→[ra:t^{ə}n]→[ra:tn]). Der Unterschied zwischen den beiden letzten Notationen sollte deshalb nicht grundsätzlich genauso bewertet werden wie der zwischen zwei deutlich wahrnehmbaren Vokalen (etwa [y] und [a]). Es scheint also so, daß bei Reliabilitätsuntersuchungen zusätzlich zu einer Qualitätsmessung (mit dem Ä-Maß) auch eine Substanzmessung (betreffend Deutlichkeit, bei synthetischer Sprache: Natürlichkeit) vorgenommen werden und in das R-Maß einfließen sollte. (4) Dies führt aber auf eine grundsätzliche Überlegung: Eignen sich Ä-Maße überhaupt als Basis für R-Maße? Ganz deutlich stellt sich die Frage, wenn aufgrund von "semantischem Hören" (an undeutlichen Textstellen) ganz andere Wörter gehört werden (vgl. z.B. /13/). An solchen Stellen können Ä-Maße nicht viel aussagen, was bedeutet, daß sich Ä-Maße als Basis von R-Maßen nur dann eignen, wenn die jeweiligen Transkripte nicht zu sehr voneinander abweichen. (5) Eines der wesentlichen Probleme bei den bekannten R-Maßen betrifft die Nicht-Beachtung möglicher systematischer Abweichungen. So können gerade bei automatischer Anwendung der Meßvorschrift systematisch auftretende Abweichungen zu scheinbar geringen Reliabilitätswerten führen, obwohl sich vielleicht durch eine einfache Translation des Referenznetzes eines Transkribenten die Reliabilität entscheidend vergrößern würde. So könnte z.B. die Aspiration eines Plosivs, etwa [p], welche ein Transkribent (T2) als [p^{h}] (bzw. [p']) notiert, bei einem anderen (T3) durchweg als [p'](bzw. [p]) erscheinen (vgl. /1/). Durch eine Transformation (Justierung) der Transkripte, also z.B. durch eine Anhebung des Aspirationsgrades bei T3 ließe sich die Reliabilität erhöhen. Wenn aber aus einer höheren Reliabilität auch eine höhere Validität folgt, stellt dieses Beispiel die deterministische Auffassung von Transkriptionsergebnissen infrage. Denn wenn nach eingehender Diskussion das [p'] des einen Transkribenten ein [p^{h}] beim zweiten bleibt, so ist das gemeinsame, justierte Transkript zwar reliabel, aber nicht mehr eindeutig. Diese Tatsache richtet den Blick direkt auf eine probabilistische Transkriptionsauffassung, die für einen Laut ggfs. mehrere Symbolalternativen, nach ihrer Wahrscheinlichkeit geordnet, zuläßt. Ein solches Transkript entspricht im übrigen auch dem "natürlichen Ergebnisverhalten" automatischer Prozesse, wo im Laufe der Berechnung immer mehrere Alternativlösungen vorhanden sind, und schließlich die wahrscheinlichste das Endergebnis bildet.

VERWENDBARKEIT FÜR DIE AUTOMATISCHE TRANSKRIPTION

Grundsätzlich lassen sich natürlich alle erwähnten Ä- und R-Maße auch zur Beurteilung automatischer Meßprozesse verwenden. Welche sind jedoch zu bevorzugen?
Typischerweise variieren akustische Meßgrößen, die die Ausgangsbasis für jeden aT-Prozeß bilden, kontinuierlich. Diesem entspricht am ehesten ein R-Maß mit einem zugrundeliegenden Ä-Maß auf Intervallskalenniveau. Gleichfalls wird man ein Ä-Maß bevorzugen, das auf akustischer bzw. auditiver Ähnlichkeit beruht. Denn die akustischen Meßgrößen (Parameter) lassen sich oft nur schwerlich mithilfe der artikulatorischen Klassifikationsdimensionen interpretieren. So kann auf der Basis akustischer Meßgrößen kaum erklärt werden, warum etwa der Abstand zwischen [p] und [t] kleiner ist als der zwischen [p] und [k], was jedes Ä-Maß auf artikulatorischer Basis grundsätzlich vorsieht. Andererseits besteht bei den auditiven Ä-Maßen aufgrund ihrer Konstruktion mithilfe von Hörtests die Gefahr einer sprachspezifischen Färbung.
Wenn das TR-Maß eines automatisch erstellten Transkripts nicht zu sehr von dem einer menschlichen Vergleichsgruppe abweicht, so weist dies auf die Brauchbarkeit des aT-Verfahrens hin. Diese Aussage läßt sich beim Übergang von den globalen TR-Maßen zu den LR-Maßen weiter verfeinern. Es ist durchaus vorstellbar, daß ein aT-Algorithmus für gewisse Lautklassen bereits akzeptable Ergebnisse liefert, für andere dagegen noch nicht. Nach Anwendung der aT müßte dann der Text abschließend nur noch einmal für kritische Lautklassen manuell überprüft werden. Der Übergang zu einem SR-Maß erlaubt schließlich die differenzierteste Beurteilung. Die Prüfung des automatischen Transkripts könnte beispielsweise nur an den maximal reliablen Textstellen stattfinden oder, wenn der Algorithmus mehrere gewichtete Alternativen ausgibt, auch diese einbeziehen.

AUSBLICK

Sind auf der einen Seite R-Maße die Voraussetzung zur Güteprüfung von rechnergestützten Analyseprozeduren, so gestatten rechnergestützte Syntheseprozeduren in Zukunft vielleicht sogar eine Validitätsprüfung. Bereits heute ist es nämlich möglich, ausgehend von einer phonetischen Symbolfolge durch Simulation des menschlichen Spracherzeugungsmechanismus synthetische Sprache zu erzeugen, die jedoch noch nicht die Qualität natürlicher menschlicher Sprache erreicht. Sollte es aber in Zukunft gelingen, den Sprechvorgang so gut nachzubilden, daß sich das Resultat weder auditiv (Hörtests!) noch artikulatorisch (Sehtests!) von einem natürlichsprachlichen unterscheidet, so wäre damit die Grundlage für eine echte Validitätsprüfung geschaffen. Die richtige Symbolfolge als Basis für die Synthese ist dann unabhängig von der Analyse (Transkription) bekannt!

LITERATUR

/1/ R. Greisbach: Grundlagen der Automatisierbarkeit phonetischer Transkription. Diss. Köln 1986 (im Druck).
/2/ P. Ladefoged: The nature of vowel quality. Rev. Lab.Fonet.Exper., Coimbra, 5 (1960) 73-162.
/3/ J. Laver: Variability in vowel perception. Lang.&Speech 8 (1965) 95-121.
/4/ G. Heike: Auditive und akustische Beschreibung lautlicher Äußerungen mit Hilfe eines lautlichen Bezugssystems. Z.f. Mundartforschung, Beih., N.F. 3 u. 4 (1967) 356-362.
/5/ W.H. Vieregge et al.: A distinctive feature based system for the evaluation of segmental description in Dutch. Proc. 10th Int. Congr. Phon.Scie. Dordrecht 1984, 654 -659.
/6/ W.H. Vieregge: Ein Maß zur Reliabilitätsbestimmung phonetisch-segmenteller Transkriptionen. Z.f. Dialektol.u. Ling. 52 (1985) 167-180.
/7/ W.H. Vieregge: The problem of validity of segmental transkriptions. Proc. Inst. Phonetics, Cath. Univ. Nijmegen 10 (1986) 23-26.
/8/ A. Almeida, A. Braun: "Richtig" und "Falsch" in der phonetischen Transkription. Z.F. Dialektol.u. Ling. 53(1986) 158-172.
/9/ B. Schädel: Über Schwankungen und Fehlergrenzen beim phonetischen Notieren. Bull. de Dial.Rom. 2 (1910) 1-29.
/10/ M. Bürkle: Zur Validität eines Maßes zur Reliabilitätsbestimmung phonetisch-segmenteller Transkription. Z.f. Dialektol. u. Ling. 53 (1986) 173-181.
/11/ J. Asendorf, H.G. Wallbot: Maße der Beobachterübereinstimmung. Z.f. Sozialpsychol. 10 (1979) 243-252.
/12/ S. Geršić: Mathematisch-statistische Untersuchungen zur phonetischen Variabilität am Beispiel von Mundartaufnahmen aus der Batschka. Göppingen 1971.
/13/ P. Winkler: Anwendungen phonetischer Methoden für die Analyse von Face-to-Face-Situationen. In: P. Winkler (Hrsg.): Methoden der Analyse von Face-to-Face-Situationen. Stuttgart 1981. 9-46.

AUTOMATIC ISOLATION OF NASAL MURMURS

H.J. WARKENTYNE

Department of Linguistics
University of Victoria
Victoria, B.C., Canada V8N 2Y2

B.C. DICKSON

Centre for Speech Technology Research
University of Victoria
Victoria, B.C., Canada V8N 2Y2

ABSTRACT

In order to determine the precise parametric values required to locate nasal murmurs in the speech signal, three routines were developed. An energy curve location routine was designed to isolate potential nasal murmurs from the data. A spectral profile-matching routine and a routine for calculating the centroid of spectral energy were then applied to the segments isolated by the energy curve location function. These operations succeeded in locating an average of 52 of 78 possible nasal murmurs for each of the ten subjects.

INTRODUCTION

The experiment reported in this paper represents a component of a research project in speaker recognition. The object was to develop a system that incorporates an automatic procedure for extracting segments from the speech signal which belong to the same phonetic class. The English nasals [m, n, ŋ] were selected as our first target since murmurs have often been shown to be significant in speaker recognition; e.g., [1], [2], [3] and [4].

The data set consisted of 88 short sentences produced by ten subjects. Each sentence contained a nasal phoneme. The phonetic context of the nasal phonemes was varied from sentence to sentence to create a wide range of environmental conditioning factors including ten vocalic environments and utterance-initial and -final positions.

Our early observations indicated that inter-speaker differences in the spectra of the nasal murmurs was a problem to be overcome before a speaker-independent nasal murmur extraction routine could be formulated. For an individual subject, detailed characterization of his nasal murmurs was required to separate them from the non-nasal segments, but this detail failed to characterize the nasal murmurs of a second subject. A series of robust parameters that isolated the nasal murmurs of all the speakers, yet did not falsely reject some murmurs on the basis of too narrow specifications, was therefore required.

As Fujimara [5] has shown, nasal murmurs can be defined in terms of three general acoustic properties independent of place of articulation, phonetic context, or individual speaker. These are: the existence of a low-frequency first formant around 300 Hz that is well isolated from any formant above it, relatively high damping factors of upper formants, and the high density or number of formants in the frequency domain, including the presence of anti-formants. The high density of weak formants should occur in the range of 300 to 2300 Hz.

Mermelstein [6] applied the above description to the development of an automatic nasal detection system for use with continuous speech recognition. He extracted four acoustic parameters using digital filtering to define four frequency bands at 0-1, 1-2, and 2-5 KHz, and a frequency centroid below 500 Hz. Digital spectra and relative intensity between frames were computed every 12.8 ms. The dynamic transition from a nasal to a vowel, or the reverse, defined by a rapid shift in the intensity, signified the probability of occurrence of a nasal murmur. The relative distribution of energy within the three frequency bands, the presence of a centroid below 500 Hz, and the dynamic shift served to indicate the presence of a nasal segment.

Our observations revealed that there was a tendency for nasal segments to be poorly defined acoustically if the duration of the murmur was less than 60 ms. It was seen that in utterance-initial environments, nasal segments were rarely accompanied by a murmur, and in utterance-final positions, a reduction in signal energy sometimes caused the murmur to be weak and irregular. Many of the phonemic nasals produced by the ten subjects appeared to be realized only on the basis of acoustic information that was a product of a transition to or from the neighbouring vowel. Because of the limited information supplied by the nasal segments in utterance-initial positions, the decision was therefore made to concentrate on isolating nasal murmurs that occur in the environment of a preceding vowel with a duration of at least 60 ms.

EXPERIMENTAL PROCEDURES

To determine the precise parametric values required to locate the nasal murmurs, three routines were developed as investigative tools on the IBM main frame. These were an energy curve location routine, a spectral profile-matching routine, and a routine for calculating the centroid of spectral energy. A high degree of flexibility was included to allow fine-tuning of values before they were incorporated into a final segment extracting system.

Energy Curve Location

To locate positions in the speech signal where overall energy of the signal dropped and maintained a steady level, a routine was first developed to convert the time-series data to an energy representation. Calculation of the signal energy was performed by passing the time series through a rectangular window and computing the mean of the squared values in the measurement interval N. The time-varying energy calculation $E(n)$ is defined by the following function:

$$E(n) = \frac{1}{N} \sum_{m=0}^{n=1} x^2(n+m)$$

where N is the number of sample points in the window.

To reduce the high amplitude of the lower frequencies in the signal, pre-emphasis was applied before energy calculation. Pre-emphasis was found to accentuate the sharp drop-off from the vowel to the nasal murmur and smooth the steady state in the energy calculation of the murmur. When pre-emphasis was applied to the signal, a greater percentage of energy in the mid-frequency range of the vowels than was present in the neighbouring nasal murmurs caused the total energy in the vowels to be accentuated.

Experimentation with the window lengths of N in the calculation of the signal energy revealed that a 20 ms window (N = 200 points, sampled at 10 K per second) yields an energy curve that is not affected by the time-varying amplitude properties of the speech signal. However, to locate the onset of the nasal murmur accurately, $E(n)$ was computed at 5 ms intervals. Thus, energy values calculated from 20 ms of time-series data were computed, advancing along the time series in 5 ms jumps.

Determination of a significant change in energy was performed in terms of ratio. For a vowel-nasal sequence, the ratio between the energy value of the triggering frame in the vowel and the lower energy value of the nasal murmur (the trigger ratio) occurring 20 ms later was found to be 1.9:1 or greater.

The following procedure was incorporated. Sequential examination of pairs of energy values representing non-overlapping 20 ms sections of the time-series data is carried out every 5 ms to locate a trigger ratio of 1.9:1 or greater. The first computed value is compared with the fifth, the second with the sixth, etc., and the ratio between the pairs is calculated. Once a trigger ratio is found, the next three consecutive pairs are examined and the pairs with the greatest ratio are selected. The high-energy value represents a 20 ms section of the time series that is the triggering frame of a potential vocalic segment. The low-energy value paired with the triggering frame represents a 20 ms portion that is potentially the start of a nasal murmur. The time co-ordinate of the low-energy frame is the start time of the steady energy level, which can be determined with an accuracy of ± 2.5ms, or half the duration of the 5 ms advance used in calculating the energy of the signal.

The steady energy level of the triggered section of the time series is also calculated by sequential examination of non-overlapping pairs of energy frames, at 5 ms intervals. The pairs must not exceed a ratio of 2:1 or fall below 0.5:1. That is, the energy value of the first frame in the steady state must not be more than double, or less than half, the energy value of the fifth frame, the same being true for the second and sixth frames, etc. The steady state must last for a minimum duration of 60 ms to be accepted. When a ratio above 2:1 or below 0.5:1 is encountered before 60 ms have elapsed, the segment is rejected. After 60 ms, these ratios are used as a shut-off and the segment of time-series data is accepted. In order to avoid acceptance of segments with gradual increases of the energy level, the segment is rejected if the value exceeded that of the triggering frame. Shut-off also occurs if the value of any frame drops below a specified value.

Profile Matching

To characterize the spectral distribution of energy common to nasal murmurs, a profile-matching routine was developed for use on the main frame. The procedure used was to select from the time-series data those sections that were isolated by energy curve location, and to create power spectra of the sections, using 50 Hz resolution and a 20 ms Hamming window advancing along the time series at 20 ms intervals. Pre-emphasis was applied to the time series. The spectra were saved on computer tape, and were later retrieved for comparison with an adjustable profile table.

The parameters incorporated for profile-matching were minimum segment duration, minimum total energy, and percentage and tolerance in up to 20 frequency buckets. The frequency buckets were defined by their upper frequency range, total percentage of all the buckets being of course 100. The routine was designed to call up the profile table upon the initiation of each operation, examine the spectrum file called up from storage, and send the results of the profile-matching to the main frame's printer. Results reported were frames matched, error vectors, and distance as a measure of closeness of fit.

As noted above, the nasal murmurs commonly show a dominance of energy in the 0-500 Hz range. To avoid the influence of individual speaker characteristics, only two frequency buckets were employed in the profile-matching routine. The first was 0-500 Hz, in which the minimum allowable percentage of energy was found to be 57% and the maximum was 99%. In the profile table this was stated as 78% of the spectral energy with a tolerance of 21%. The second frequency bucket was 500 to 5000 Hz, in which the remainder of the energy in the spectrum could be distributed. This was stated as 22% with a tolerance of 21%.

Spectral Centroid Determination

The frequency centroid of a spectrum is essentially the mean frequency of energy in the power spectrum, and is determined by the formula

$$\text{Centroid frequency} = \frac{\sum_{i=1}^{n} f_i I_i}{\sum_{i=1}^{n} I_i}$$

where f is the frequency of bin i, I is the intensity of bin i, and n is the number of the last bin that corresponds to a frequency not greater than the cut-off frequency defined for the centroid calculation. A Fortran program was written to perform this calculation, using as input the power spectra held in storage on the main frame. An adjustment to the upper frequency *n* was included so that the centroid could be determined for any lowpass bandwidth of the power spectrum. The results of the calculations were displayed in the time domain.

When the centroid calculation was applied to the full 5 KHz passband, a large number of non-nasal segments had a frequency centroid in the range of nasal murmurs; i.e., below 600 Hz. These segments included [ɹ, l, w] voiced stops, [u] and unstressed [i] and [ə], especially when the consonants combined with the vowels. Reducing the cut-off frequency to 1000 Hz eliminated most of the above unwanted segments if only segments with a centroid below 400 Hz were accepted. Although some of the non-nasal elements still exhibited low centroids, the 60 ms duration criterion succeeded in eliminating a majority of these.

Sequencing of Routines

The energy curve location routine was designed to scan the time-series data to determine areas in the speech signal where a nasal murmur was possible. Since this did not require conversion to Fourier series, it was the most economical of the routines to apply to the full data to isolate potential nasal segments. These could then be converted to the Fourier series and be processed by the profile-matching and centroid routines. The latter two routines were applied independently of each other and therefore did not require a particular order.

RESULTS

The speech data of the ten subjects under analysis contained a total of 780 post-vocalic nasal phonemes. Of these, the energy curve location routine successfully isolated 593. The routine also isolated 655 non-nasal phonetic events or sequences of events that took place in a post-vocalic environment. A further 155 non-nasal phonetic events were captured after being triggered by a high-energy non-vocalic signal, indicating the need to incorporate a subroutine that will examine the triggering frame to determine the presence of voicing.

Of the 593 potential nasal murmurs isolated by the energy curve location, the combined profile matching and centroid locating functions accepted 516. When performed independently, the profile-matching routine rejected 356 non-nasals and the centroid calculation rejected 330. The combined effect resulted in the rejection of 454 of the 655 sections of unwanted data.

For a large group of subjects, where robust parameters must be applied in order to isolate the segments, interspeaker characteristics interfere with the process of distinguishing the nasal murmurs from the non-nasals. We have found, however, that speaker-specific characteristics may be used to describe the quality of the nasal murmurs, thereby creating a criterion for rejecting most of the non-nasals. It is apparent from our observations that speaker-specific characteristics are recurrent in the majority of the nasal murmurs. A statistical approach might therefore be usefully employed to describe automatically the mean spectral characteristics of the speech events accepted by the system. A comparison could then be made of each spectral series to determine its closeness of fit to the mean, and, using a degree of tolerance or a distance metric, deviant spectral data could be rejected.

REFERENCES

[1] J.W. Glenn and N. Kleiner, "Speaker identification based on nasal phonation ", *Journal of the Acoustical Society of America 43*: 368-372, 1968.

[2] M.R. Sambur, "Selection of acoustic features for speaker identification", *IEEE Transactions in Acoustics, Speech and Signal Processing ASSP-23*: 169-176, 1975.

[3] L-S. Su, K-P. Li, and K.S. Fu, "Identification of speakers by use of nasal coarticulation", *Journal of the Acoustical Society of America 56*: 1876-1882, 1974.

[4] J.J. Wolf, "Efficient acoustic parameters for speaker recognition", *Journal of the Acoustical Society of America 51*: 2044-2056, 1972.

[5] Osamu Fujimara, "Analysis of nasal consonants", *Journal of the Acoustical Society of America 34*: 1865-1875, 1962.

[6] P. Mermelstein, "On detecting nasals in continuous speech", *Journal of the Acoustical Society of America 61*: 581-587, 1977.

A SEMIVOWEL RECOGNITION SYSTEM*

Carol Y. Espy-Wilson

Department of Electrical Engineering and Computer Science
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Abstract

We discuss a framework for an acoustic-phonetic approach to speech recognition. The recognition task is the class of sounds known as the semivowels (w,l,r,y) and the results obtained across several data bases are fairly consistent. We discuss some issues which were manifested by this work. These issues include feature spreading, the assignment of phonetic labels and lexical representation.

Introduction

We have developed a framework for an acoustic-phonetic approach to speech recognition. Such an approach consists of four basic steps. First, the features needed to recognize the sound(s) of interest must be specified. Second, acoustic correlates of the features must be determined. Third, algorithms to extract the properties must be developed. Finally, the properties must be integrated for recognition.

In this paper, we discuss briefly the application of the above mentioned steps to the development of a recognizer of voiced and nonsyllabic semivowels of American English. In addition, we discuss some issues brought forth by this work. These issues include feature spreading and how it can possibly be explained with a theory of syllable structure, how feature spreading affects lexical access, and if and when phonetic labels should be assigned to acoustic events.

Corpora

The initial step in this research was the design of a data base for developing and testing the recognition algorithms. We chose 233 polysyllabic words from the 20,000 word Merriam Webster Pocket dictionary. These words contain the semivowels and other similar sounds in many different contexts. The semivowels occur in clusters with voiced and unvoiced consonants and they occur in word initial, word final and intervocalic positions. The semivowels are also adjacent to vowels which are stressed and unstressed, high and low, and front and back.

For developing the recognition algorithms, the data base was recorded by two males and two females. We refer to this corpus as Database-1. Two corpora were used to test the recognition system. Database-2 consisted of the same polysyllabic words spoken by two new speakers, one male and one female. Database-3 consisted of a small subset of the sentences in the TI data base [1]. In particular, we chose two sentences which contained a number of semivowels. One sentence was said by 6 females and 8 males. The other sentence was said by 7 females and 8 males. The speakers covered 8 dialects.

Several tools described in [2] were used in the transcription and analysis of the data bases. Database-1 and Database-2 were transcribed by the author and Database-3 was segmented and labelled by several experienced transcribers.

Features, Properties and Parameters

To recognize the semivowels, features are needed for separating the semivowels as a class from other sounds and for distinguishing between the semivowels. Shown in Tables 1 and 2 are the features needed to make these classifications. The features listed are modifications of ones proposed by Jakobson, Fant and Halle [3] and by Chomsky and Halle [4]. In the tables, a "+" means that the speech sound(s) indicated has the designated feature and a "−" means the speech sound(s) does not have the designated feature. If there is no entry, then the feature is not specified or is not relevant.

An acoustic study [5] was carried out in order to supplement data in the literature (e.g., [6]) to determine acoustic correlates for the features. The mapping between features and acoustic properties and the parameters used in this process are shown in Table 3. As indicated, no absolute thresholds are used to extract the properties. Instead, we used relative measures which tend to make them independent of speaker, speaking rate and speaking level. The properties are of two types. First, there are properties which examine an attribute in one speech frame relative to another speech frame. For example, the property used to capture the nonsyllabic feature looks for a drop in either of two mid-frequency energies with respect to surrounding energy maxima. Second, there are properties which, within a given speech frame, examine one part of the spectrum in relation to another. For example, the property used to capture the features front and back measures the difference between F2 and F1.

To quantify the properties, we used a framework, motivated by fuzzy set theory [7], which assigns a value within the range

	voiced	sonorant	nonsyllabic	nasal
voiced fricatives,stops,affricates	+	−	+	−
unvoiced fricatives,stops,affricates	−	−	+	−
semivowels	+	+	+	−
nasals	+	+	+	+
vowels	+	+	−	−

Table 1: Features which characterize various classes of consonants

*Supported by a Xerox Fellowship

	stop	high	back	front	labial	retroflex
/w/	−	+	+	−	+	−
/y/	−	+	−	+	−	−
/r/	−	−		−	−	+
light /l/	+	−		−	−	−
dark /l/	−	−	+	−	−	−

Table 2: Features for discriminating between the semivowels

Feature	Acoustic Correlate	Parameter	Property
Voiced	Low Frequency Periodicity	Energy 200-700 Hz	High*
Sonorant	Comparable Low & High Frequency Energy	Energy Ratio $\frac{(0-300)}{(3700-7000)}$	High
Nonsyllabic	Dip in Energy	Energy 640-2800 Hz	Low*
		Energy 2000-3000 Hz	Low*
Stop	Abrupt Spectral Change	1st Difference of Bandlimited Energies (positive & negative)	High
High	Low F1 Frequency	F1 − F0	Low
Back	Low F2 Frequency	F2 − F1	Low
Front	High F2 Frequency	F2 − F1	High
Labial	Downward Transitions for F2 and F3	F3 − F0	Low*
		F2 − F0	Low*
Retroflex	Low F3 Frequency & Close F2 and F3	F3 − F0	Low
		F3 − F2	Low

Table 3: Parameters and Properties

*Relative to a maximum value

[0,1]. A value of 1 means we are confident that the property is present, while a value of 0 means we are confident that it is absent. Values between these extremes represent a fuzzy area indicating our level of certainty that the property is present/absent.

Control Strategy

Phonotactic constraints are used heavily in the recognition system. These constraints state that semivowels almost always occur adjacent to a vowel. Therefore, they are usually prevocalic, intervocalic or postvocalic. For recognition, these contexts map into three types of places within a voiced sonorant region. First the semivowels can be at the beginning of a voiced sonorant region, in which case they are prevocalic. Second, the semivowels can be at the end of a voiced sonorant region, in which case they are postvocalic. Finally, the semivowels may be further inside a voiced sonorant region. We refer to these semivowels as intersonorant, and one or more may be present within such a region. Semivowels of this type can be either intervocalic or in a cluster with another sonorant consonant such as the /y/ in "banyan." Although there is one overall recognition strategy, there are modifications for these contexts.

The recognition strategy for the semivowels is divided into two steps: detection and classification. The detection process marks certain acoustic events in the vicinity of times where there is a potential influence of a semivowel. In particular, we look for minima in the mid-frequency energies and we look for minima and maxima in the tracks of F2 and F3. Such events should correspond to some of the features listed in Tables 1 and 2. For example, an F2 minimum indicates a sound which is more "back" than an adjacent segment(s). Thus, this acoustic event will occur within most /w/'s and within some /l/'s and /r/'s.

Once all acoustic events have been marked, the classification process integrates them, extracts the needed acoustic properties, and through explicit semivowel rules decides whether the detected sound is a semivowel and, if so, which semivowel it is. An example of this process is illustrated with the word "flourish" shown in Figure 1.

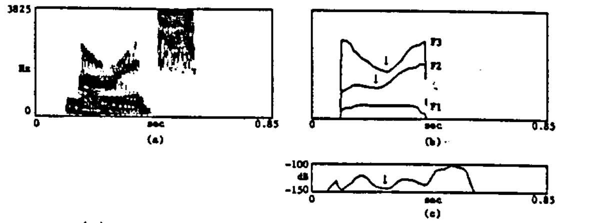


Figure 1: (a) Spectrogram of the word "flourish," (b) formant tracks and (c) Energy 640 Hz to 2800 Hz.

As can be seen, several acoustic events signal the presence of the intervocalic /r/. These events include an energy dip, a small F2 dip and a strong F3 dip. Given the energy dip marked in part c, the recognition system will extract the surrounding energy maxima corresponding to syllabic nuclei. These latter points are used to define a region for further analysis of the detected sound. Among the various events, the F3 dip is the most prominent one which gives some clue to the identity of the detected sound. Thus, it is in a small region surrounding the time of this event that the formant based properties are extracted. In addition, it is between the time of the F3 dip and the surrounding energy peaks that we characterize the rate of spectral change to determine its degree of abruptness.

Once the properties listed in Table 3 are extracted for the detected sound, the control strategy, on the basis of the types of events marked, decides which semivowel rules to apply. Again, since there is a strong F3 dip, the /r/ rule is applied first. The only other semivowel which is expected to sometimes have a sizeable F3 dip is the labial sound /w/. Thus, the /w/ rule is applied if the /r/ rule receives a low score (< 0.5).

Rules for integrating the properties were written for each of the semivowels. In addition, because they are acoustically similar, a rule was written for identifying a class that could be either /w/ or /l/. Across contexts, the rules are similar. However, well known acoustic differences between allophones such as the closer spacing between F2 and F1 for sonorant-final /l/'s as opposed to sonorant-initial /l/'s are accounted for. Additionally, within the rules, primary versus secondary cues are distinguished. For example, the /r/ rule states that if the detected sound is retroflexed, classify it as an /r/. However, if the sound is "maybe" retroflexed, look at other cues before making a decision.

Since the value of each property lies between 0 and 1, the score of any rule within the fuzzy logic framework is also in this range. Thus, we consider a sound to be classed as a semivowel if the result of a rule is greater than or equal to 0.5.

Recognition Results

The overall recognition results are given in Table 4 for each of the data bases. The term "nc" in the table means that one or more semivowel rules was applied, but the score(s) was less than 0.5. The term "others" refers to flaps, voiced /h/'s and sonorant-like voiced consonants.

As can be seen, there is quite a bit of confusion between /w/ and /l/. However, the degree to which they are confused varies considerably with context. For example, when they are prevocalic and are not preceded by a consonant, the system correctly classifies 80% of the /w/'s in Database-1 and 67% of the /w/'s in Database-2. Likewise, it correctly classifies 63% of the /l/'s in Database-1 and 76% of the /l/'s in Database-2. This context is not covered in Database-3. However, 71% of the prevocalic /w/'s adjacent to unvoiced consonants in Database-3 were classified correctly. Considering the many differences between Database-3 and the other corpora which include coverage of contexts, coverage of dialects, recording methods and transcription biases, the results across data bases are quite consistent.

	w	l	r	y	nasals	others	vowels	
# tokens	369	540	558	222	464	508	2385	
undetected(%)	1.4	3.3	2.6	2.9	24	81.5		
w(%)	52	7.5	3.4	0	1	1	1	
l(%)	9.1	55.7	0	0	11	3.3	5.5	
w-l(%)	31.4	30.4	0	0	3	.8	2	Database-1
r(%)	4	.2	90	0	2	.6	6	
y(%)	0	0	0	93.7	6	1.4	8.6	
nc(%)	2	3	4.7	4.9	53	11.4	39	
# tokens	181	274	279	105	232	135	1184	
undetected(%)	1.7	1.5	4.3	3.8	24	69		
w(%)	48	3.6	1.9	0	5	0	1	
l(%)	12.7	57.7	0	0	7	6	5	
w-l(%)	29	33.8	0	0	3	1	4	Database-2
r(%)	3.5	.4	91.3	0	3	2	4	
y(%)	0	0	0	84.9	3	3	10	
nc(%)	6.7	2.9	4.3	13.3	55	19	42	
# tokens	28	40	49	23	44	121	350	
undetected(%)	3.6	7.5	0	4	50	73		
w(%)	46	10	0	0	15	0	2	
l(%)	21.6	52.6	0	0	13	2.5	9	
w-l(%)	21.6	24.7	0	0	0	0	4	Database-3
r(%)	7.1	0	89.8	0	5	2.5	15	
y(%)	0	0	0	78.5	0	5	9	
n-c(%)	0	5.1	10.2	17.2	17	17	62	

Table 4: Overall Recognition Results

From Table 4 we see that there are several "misclassifications" of nasals, vowels and other sounds as semivowels. It is important to note, however, that the system has no method for detecting the feature "nasalization." Therefore, the distinction between nasals and semivowels lies mainly in the abruptness of spectral change surrounding the detected sounds. As in the case of the nasals, some misclassifications of vowels and other sounds as semivowels can be eliminated by including other features in the recognition system and by refining the parameters. However, the avoidance of other confusions is not straightforward (In addition, some of the misclassifications do not appear to be errors of the system, but errors in the transcription). It is this issue which is addressed in the remainder of the paper.

Discussion

This research has highlighted several interrelated issues which are important to any recognition system based on an acoustic-phonetic approach. One such issue relates to the spreading of one or more features of a sound to a nearby segment, thereby resulting in a change of some of the features of the segment and possibly a merging of the two segments. Although examples of this phenomenon occurred with several features, we will discuss it in the context of the feature retroflexion which appears highly susceptible to spreading. Examples are illustrated in Figure 2 with the words "cartwheel" and "harlequin."

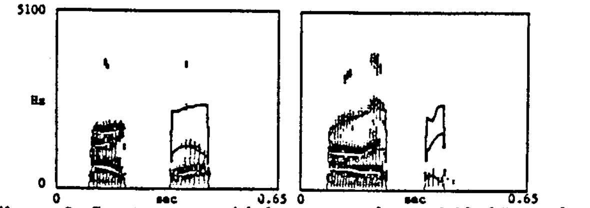


Figure 2: Spectrograms with formant tracks overlaid of "cartwheel" (left) and "harlequin" (right).

In each instance, it appears as if the underlying /r/ and adjacent vowel combine such that their acoustic realization is an r-colored vowel. The occurrence of such feature assimilation is predicted by the syllable structure theory as explained by Selkirk [8]. This syllable structure is shown in Figure 3, where the onset consists of any syllable-initial consonants, the peak consists of either a vowel or vowel and sonorant, and the coda consists of any syllable-final consonants. Selkirk states that when /l/ or /r/ is followed by a consonant which must occupy the coda position, it becomes part of the peak. Thus, the structure for the first syllable in "cartwheel" is as shown in Figure 4. Since the /ɑ/ and /r/ both occupy the syllable peak, we might expect some type of feature assimilation to occur. If it is true that a vowel and /r/ in this context will always overlap to form an r-colored vowel, then no exception is needed in the phonotactic constraints of semivowels for words like "snarl" where the /l/ is "supposedly" separated from the vowel by the /r/. Instead, the constraints can simply state that semivowels must always be adjacent to a vowel.

When a postvocalic /l/ or /r/ is not followed by a syllable-final consonant, Selkirk states that it will tend to be in the coda although it has the option of being part of the peak. This option was clearly exercised across the speakers in Database-1 and Database-2. As an example, consider the two repetitions of the word "carwash" shown in Figure 5. As in the word "harlequin," the /ɑ/ and /r/ in the word "carwash" on the left appears to be one segment in the sense that retroflexion extends over the entire vowel duration. However, in the repetition on the right, the /ɑ/ does not appear to be retroflexed. Instead, there is a clear downward movement in F3 which separates the /ɑ/ and /r/ and thus the /r/ appears to be syllable-final.

We dealt with this feature spreading phenomenon in the recognition system by considering it a correct classification if the vowels in words like "cartwheel," "harlequin" and "carwash" were labeled /r/. This seemingly "disorder" was allowed since the vowel's and following /r/'s appear completely assimilated.

Allowing this "disorder" at the acoustic level means that the ambiguity must be resolved at or before lexical access. There is at least one example in the data bases where a seemingly prevocalic /r/ and adjacent vowel merged to form an r-colored vowel. If this is so, then there does not appear to be a clear method for

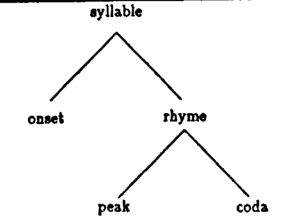


Figure 3: Tree structure of syllable.

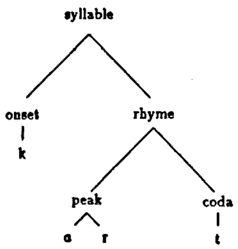


Figure 4: Tree structure of syllable "cart."

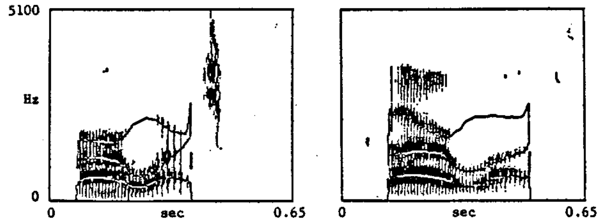


Figure 5: Spectrograms with formant tracks overlaid of two repetitions of "carwash."

determining whether an r-colored vowel is underlyingly a vowel followed by /ɹ/ or a vowel preceded by /ɹ/.

This ambiguity as well as the fact that some vowels and other voiced consonants are classified as semivowels raises the issue of whether or not phonetic labels should be assigned before lexical access. In other words, is the representation of items in our lexicon in terms of phonetic labels or features?

If we assume that lexical items consist of a sequence of phonetic labels, then it is clear from an analysis of the misclassifications made in the semivowel recognition system that context must be considered before phonetic labels are assigned. That is, some sounds are misclassified because contextual influences caused them to have patterns of features which normally correspond to a semivowel. For example, consider the word "forewarn" shown in Figure 6. Because of the labial F2 transition and the downward F3 transition arising from the adjacent /ɹ/, the beginning of the first /ɔ/ was classified as a /w/. It is clear in cases like this that if phonetic labels are going to be assigned, context should be considered before it is done. The issue then becomes, how much context needs to be considered. For example, consider the word "fibroid" also shown in Figure 6 which has a fairly steady state F3 frequency of about 1900 Hz. We have observed that in words like this where a labial consonant is preceded by a normally non-retroflexed vowel and followed by a retroflexed sound, the first vowel can be totally or partially retroflexed. Such feature spreading is not surprising when we consider that the intervening labial consonant does not require a specific placement of the tongue.

If, instead of phonetic labels, lexical items are represented as matrices of features, it may be possible to avoid misclassifications due to contextual influences and feature spreading since we are not trying to identify the individual sounds before lexical access. For example, consider the comparison given in Table 5 of what may be a partial feature matrix in the lexicon for an /ɑ/ and postvocalic /ɹ/ with property matrices for these segments in the words "carwash" shown in Figure 6. The lexical representation is in terms of binary features whereas the acoustic realizations are in terms of properties whose strengths as determined by fuzzy logic lie between 0 and 1.

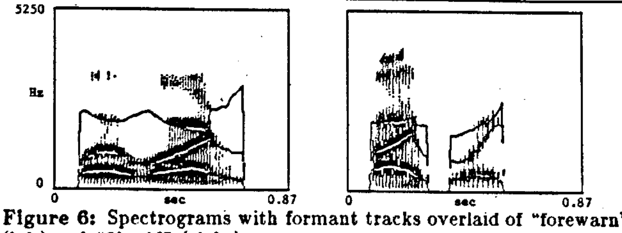


Figure 6: Spectrograms with formant tracks overlaid of "forewarn" (left) and "fibroid" (right).

	lexical representation		realisation #1		realisation #2
	ɑ	ɹ	ɑ	ɹ	$ɑ^r$
high	–	–	0	0	0
low	+	–	1	0	1
back	+	±	1	1	1
retroflex	–	+	0	1	1

Table 5: Lexical Representation vs. Acoustic Realizations of /ɑɹ/.

Acoustic realization #1 and the lexical representation are a straightforward match. (Assume a simple mapping strategy where property values less than 0.5 correspond to a "–" and property values greater than or equal to 0.5 correspond to a "+.") However, the mapping between acoustic realization #2 and the lexical representation is not as obvious. It may be possible for a metric to compare the two representations directly since the primary cues needed to recognize the /ɑ/ and /ɹ/ are unchanged. On the other hand, we may need to apply feature spreading rules before using a metric. The rules can either generate all possible acoustic manifestations from the lexical representation or generate the "unspread" lexical representation from the acoustic realization.

Determining the mapping between features and properties which have varying degrees of strength is an important and difficult problem which may give insights into the structure of the lexicon. The solution to this problem will require a better understanding of feature assimilation in terms of what features are prone to spreading, and in terms of the domains over which spreading occurs. Resolution of these matters is clearly important to an acoustic-phonetic approach to speech recognition.

REFERENCES

[1] Lamel, L., Kassel, R., and Seneff, S., "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. Speech Recog. Workshop*, CA., 1986.

[2] Cyphers, D., Kassel, R., Kaufman, D. Leung, H., Randolph, M. Seneff, S., Unverferth, J., Wilson, T. and Zue, V. "The Development of Speech Research Tools on MIT's Lisp Machine-Based Workstations," *Proc. Speech Recog. Workshop*, CA, 1986.

[3] Jakobson, R., Fant, G. and Halle, M., "Preliminaries to Speech Analysis," *MIT Acoustics Lab. Tech. Rep. No. 13*, 1952.

[4] Chomsky, N. and Halle, M. *The Sound Pattern of English*, New York: Harper and Row, 1968.

[5] Espy-Wilson, Carol Y., "An Acoustic-Phonetic Approach to Speech Recognition: Application to the Semivowels," Doctoral Dissertation, MIT, to be completed in June 1987.

[6] Lehiste, I., "Acoustic Characteristics of Selected English Consonants," *Report No. 9*, U. of Mich., Comm. Sci. Lab., 1962.

[7] DeMori, Renato, *Computer Models of Speech Using Fuzzy Algorithms*. New York: Plenum Press, 1983.

[8] Selkirk, E.O., "The Syllable," *The Structure of Phonological Representations (part II)*, ed. van der Hulst, H. and Smith N., Dordrecht: Foris Publications, 1982.

PHONEME-BY-PHONEME RECOGNITION AND SEMANTIC INTERPRETATION OF MULTI-SPEAKER SPEECH (THE HCDP-APPROACH)

TARAS VINTSYUK

Speech Recognition and Synthesis Laboratory
Institute of Cybernetics
Kiev, Ukraine, USSR 252207

ABSTRACT

A new approach to phoneme-by-phoneme recognition and semantic interpretation of multi-speaker speech is proposed. The approach is based on a constructive (C) representation of complex speech signals with hierarchic (H) structure of speech patterns (signals, microphonemes, phonemes, diphones, syllables, words, sentences, communicated senses). The recognition and semantic interpretation reside in composing (C) for a given speech signal and subsequent parsing of such complex speech pattern that is consistent with all levels of the hierarchy and is the most similar in some sense to the one to be recognized. The guided composition and subsequent parsing of this complex speech signal are realized by means of dynamic programming (DP). Some examples of solved problems are listed.

SPEECH PATTERN HIERARCHY AND MATHEMATICAL MODELS OF SEGMENTS

The HCDP-method integrates the approved principles of speech information decoding and processing and generalizes the CDP-method /1/, /2/, /3/. The hierarchic principle presumes the hierarchy of the patterns. The speech signals are discribed by sequences of observable elements-vectors x_i: $X_l = (x_1, x_2, \dots, x_i, \dots, x_l)$, where l is a length of the speech signal in uniform or quasi-uniform discrete time with spacing (mean spacing) of 15 ms for instance. The subsequences of the elements $X_{mn} = (x_{m+1}, x_{m+2}, \dots, x_n)$ being named segments are interpreted as the speech patterns or more precisely as the realisations of first-level speech patterns (the microphonemes, phonemes, diphones, or syllables). In this case X_{mn} is considered as the first-level segment. Sets of the signals X_{mn} for the first-level patterns j^1 are specified by distributions $p(X_{mn}/j^1)$, $j^1 \in J^1$, where J^1 is an alphabet of the first-level patterns. The second-level speech patterns j^2 are specified by the transcriptions in the alphabet of the first-level patterns: $j^2 = (j_1^1, j_2^1, \dots, j_s^1, \dots, j^1_{q(j^2)})$, where $q(j^2)$ is the length of the transcription of the pattern $j^2 \in J^2$, J^2 is the alphabet of the second-level patterns. The second-level segments correspond to the second-level patterns and are composed (the composition (C)) of the first-level segments by merging them into the sequences in conformity with the second-level pattern transcription. For instance if the microphonemes or phonemes are the first-level patterns then the phonemes or diphones (syllables) can be the second-level patterns correspondingly.

The patterns and segments at the next hierarchic levels (the syllables, words, sentences, communicated senses) are defined similarly. Let $j^r \in J^r$, $j^r = (j_1^{r-1}, j_2^{r-1}, \dots, j_s^{r-1}, \dots, j^{r-1}_{q(j^r)})$, $p(X_{mn}/j^r)$ be the r-level pattern from the alphabet J^r, the transcription of the pattern j^r and the probability of the r-level segment X_{mn} under the condition of the pattern j^r correspondingly. The top-level patterns in the hierarchy - the communicated sense from a given finite set of senses - are specified by a canonical form and a formal constraction being named a directed semantic network and sense types and sentence types /2/, /3/. While forwarding to publications /2/, /3/ for the details let us concentrate attention on a fact that the top-level hierarchic patterns - the communicated sense - are specified actually by a list of the sentences that express the same sense. But this specification is realized by some memory-saving means instead of direct enumeration. From this more accurate definition it also follows that the r-level pattern is not obligatory expressed with one transcription in the alphabet of the (r-1)-level patterns and there can be several or even many such transcriptions.

Constructive (C1) nature of the model manifests in expressing the probabilities of the segments X_{mn} under the condition of the r-th pattern $p(X_{mn}/j^r)$ as products of the probabilities of the corresponded to the transcription j^r segments under the condition of the patterns of the previous (r-1)-th level:

$$p(X_{mn}/j^r) = \prod_{s=1}^{q(j^r)} p(X_{m_{s-1}m_s}/j_s^{r-1}) \quad , \quad (1)$$

where m_s are bounds of the (r-1)-level segments: $m_o=m$, $m_{s-1}< m_s$, $m_{q(j^r)} = n$. Thus the probability of the observed (to be recognized) signal $X_1 = X_{o1}$ under the condition of the top-level hierarchic pattern $j^R \in J^R$, J^R is the alphabet of the top-level hierarchic patterns, takes a form of the product of the corresponding segment probabilities under the condition of the first-level patterns:

$$p(X_{o1}/j^R) = \prod_{s=1}^{\overline{q}(j^R)} p(X_{m_{s-1}m_s}/j_s^1) \quad . \quad (2)$$

In the expression (2) $\overline{q}(j^R)$ is a number of the first-level patterns from the sequence of which the top-level hierarchic pattern j^R is composed, $m_o = 0$, $m_{s-1}< m_s$, $m_{\overline{q}(j^R)} = 1$ are the bounds of the first-level segments.

To describe (specify) the mathematical model of the speech signals and to use it then for solving the speech recognition problems there is obviously sufficient to give the transcriptions of the patterns at all levels of the hierarchy with the segment distributions under the condition of all first-level hierarchic patterns $p(X_{mn}/j^1)$, $j^1 \in J^1$. These distributions are specified for every possible segment length n-m that takes generally different values for the different patterns $j^1 \in J^1$.

In line with the expression (1) it may seem that the segments of the speech signal are considered as mutually independent ones. In reality it follows just from the expression (1) as well as (2) that there is a strong deterministic dependence of the segments in the sequences that is manifested in constraints on the pattern order in the sequences, i.e. is expressed in the transcriptions of the patterns at all levels of the hierarchy.

RECOGNITION CRITERION AND METHOD

By using the maximal likelihood method let us classify the signal X_1 to be recognize as such top-level hierarchic speech pattern that the acceptable for the subject field sequence of the first-level patterns that is composed by the transcriptions in accordance with the pattern hierarchy will induce on X_1 such first-level segmentation for which the likelihood expression reaches an absolute maximum:

$$j^R(X_1) = \underset{j^R \in J^R}{\operatorname{argmax}} \max_{\{m_s\}} \prod_{s=1}^{\overline{q}(j^R)} p(X_{m_{s-1}m_s}/j_s^1). \quad (3)$$

The expression (3) presumes the exhaustive search through all pattern transcriptions if the patterns are specified not by one but by two or more transcriptions.

The recognition criterion (3) determines top-down and down-top analysis of the signal X_1 simultaneously. It is important that by solving the problem (3) one receives a consistent with all hierarchic levels interpretation referring if necessary to the segment borders of all-level hierarchic patterns being contained in the analizing signal.

By analizing the expression (3) and taking account of a fact that the borders of the r-level segments unconditionally coincide with the borders of certain (r-1)-level segments one concludes that the exhaustive search to maximize the expression (3) can be avoided and the solution can be found by the Bellman's optimality principle with help of the dynamic programming. For computation it is more convenient to use a logarithm of the likelihood. The expression (3) is transformed into

$$j^R(X_1) = \underset{j^R \in J^R}{\operatorname{argmax}} \max_{\{m_s\}} \sum_{s=1}^{\overline{q}(j^R)} \ln p(X_{m_{s-1}m_s}/j_s^1). \quad (4)$$

The constructivity (C2) of the HCDP-method is just in referencing the effective method to maximize (4) for the segment borders $\{m_s\}$ and all-level hierarchic patterns - in using the dynamic programming (DP) for these goals.

To afford the constructivity C2 one needs the constructive (C3) techniques to specify the hierarchy of the patterns and their transcriptions and the constructive (C4) means to describe the distributions $p(X_{mn}/j^1)$, $j^1 \in J^1$ under the condition of the first-level patterns for every possible segment length. Let us consider the realization of the constructivity principles with the particular examples from /2/, /3/.

MICROPHONEMIC RECOGNITION AND SEMANTIC INTERPRETATION

The first level of the hierarchy is the microphonemes (parts of the phonemes). The microphoneme j^1 is specified by one or more standard elements being denoted by $e(j^1)$ and having more frequently the same physical nature as the observed speech elements. The distribution of the segment X_{mn} under the condition of the microphoneme j^1 is defined by the relationship:

$$G(X_{mn}, j^1) = \ln p(X_{mn}/j^1) = \sum_{i=m+1}^{n} \ln p(x_i/e(j^1)) = \sum_{i=m+1}^{n} g(x_i, e(j^1)), \quad (5)$$

where the segment length satisfies the condition

$$T_{min}(j^1) \leqslant n-m \leqslant T_{max}(j^1) \quad . \quad (6)$$

In accordance with (5)-(6) one considers the quantity $g(x_i, e(j^1))$ as an elementary measure of similarity between the observed element x_i and standard element $e(j^1)$, and $G(X_{mn}, j^1)$ as the similarity between the segment X_{mn} and the first-level pattern j^1 such that the latter itself is the stationary segment being composed of one element $e(j^1)$ that is replicated n-m times to quote the constraints (6). The number of the microphonemes $j^1 \in J^1$ is 128, 256, 512, but not greater than 1024.

The second level of the hierarchy is the words that are specified by one or more so-called acoustic or Q-transcriptions - the sequences that are composed of the first-level patterns /2/, /3/.

The third level of the hierarchy is the arbitrary word sequences being composed of the free-ordered words from a selected vocabulary. The fourth level is the allowable sentences of the subject field that are specified by the sentence types, or sence types, or directed semantic network /2/, /3/. The fifth level is a canonic form of the communicated sense.

By restricting to the first two or three levels a system is obtained to recognize correspondingly the words or continuous speech that is composed of the words from the chosen vocabulary.

PHONEME-BY-PHONEME (DIPHONIC) RECOGNITION AND SEMANTIC INTERPRETATION

The diphonic model of speech signal generation /2/, /3/ is a good compromise reflecting dynamic properties of the speech signals and realizing the phonemeness principle in the recognition. Let us insert in the hierarchic model being dealt in the previous section an additional level - the level of the diphones that takes an intermediate place between the level of the microphonemes and the level of the words. The diphonic word transcriptions are evidently defined by their phonetic transcriptions in a unique manner. The obtained six-level speech recognition and semantic interpretation system is realized the phoneme-by-phoneme recognition principle more evidently.

ZERO LEVEL OF THE HIERARCHY - MULTIDIMENSIONAL (VECTOR) QUANTIZATION

The constructivity (C5) of the HCDP-method is in using the principle of the vector quantization of the speech signals, i.e. in inserting the zero-level hierarchic patterns where the observed sequences $X_l=(x_1, x_2,\dots,x_i,\dots,x_l)$ from the vectors-elements x_i are replaced by the sequences $I_l=(j_1^o, j_2^o,\dots,j_i^o,\dots,j_l^o)$ from vectors-scalars $j_i^o = j^o(x_i)$: each observed element-vector x_i is replaced by a number of a domain $j_i^o = j^o(x_i)$ to which the observed element x_i belongs in the multidimensional space of the signals x, $j^o \in J^o$, where J^o is the alphabet of the zero-level patterns.

The introduction of the zero-level patterns allows to go over from an investigation of the relationships in the vector sequences to the investigation of the relationships in the sequences of the scalars. Now one ought to substitute the distributions $p(I_{mn}/j^1)$, $j^1 \in J^1$, where $I_{mn}=(j^o_{m+1}, j^o_{m+2},\dots, j^o_n)$ for the distributions $p(X_{mn}/j^1)$ in the formulas (1)-(5). Then in line with the principle C4 one should point out the constructive principles of specifying the distribitions $p(I_{mn}/j^1)$ for the allowable values of n-m. The first group is the methods based on a tabular specification of the distributions $p(I_{mn}/j^1)$, on an effective storing these distributions in the networks, or simply on storing the encountered values I_{mn} under the condition of the pattern $j^1 \in J^1$. In the second group there are the methods based on an approximation of the distributions $p(I_{mn}/j^1)$ with help of simple expressions and on usage of the formulas that are analogous with (5). One example:

$$p(I_{mn}/j^1)=\prod_{i=m+1}^{n} p(j_i^o/j^1) \text{ or } G(I_{mn},j^1)=\sum_{i=m+1}^{n} \ln p(j_i^o/j^1).$$

Here the distributions $p(I_{mn}/j^1)$ are specified by the tables of $|J^o| * |J^1|$ numbers $p(j^o/j^1)$.

LEARNING TO RECOGNITION AND MULTI-SPEAKERNESS

The necessary knowledge base - such a priori data as the pattern hierarchy, subject field, syntax, semantics, vocabulary, alphabets and transcriptions of the upper-level patterns - is prepared beforehand by a creator of the speech signal recognition systems. The remained undefined data (the alphabets of the lower-level patterns, the corresponding transcriptions of the lower-level patterns, the distributions $p(X_{mn}/j^1)$ or $p(I_{mn}/j^1)$ for all first-level hierarchic patterns) are computed in a learning-to-recognition mode from a multi-speaker learning set.

References

/1/ T.K.Vintsiuk, CPD-methodes de reconnaissance et d'interpretation de la parole, "Le Symposium Sovietico-Francais sur "Le Dialogue Acoustique de l'Homme avec la Machine", Moscou, 1984, p. 38 - 41.

/2/ T.K.Vintsiuk, Speech recognition and semantic interpretation, "Kibernetika", 1982, No.5, p. 101 - 111 (in Russian).

/3/ T.K.Vintsiuk, Analysis, recognition and interpretation of speech signals, Kiev, "Naukova Dumka", 1987, 280 p. (in Russian).

THE LIMITS OF SEGMENTAL DESCRIPTION

FRANCIS NOLAN

Department of Linguistics
Cambridge University
Cambridge CB3 9DA, UK

ABSTRACT

Evidence is discussed which perturbs the segmental, categorial foundation of descriptive phonetics. EPG studies showed that in cases which would be treated in auditory phonetic analysis and in phonological description as place assimilation, there is often a residual gesture towards the 'underlying' segment. Such results underline that the performance of segmental contrasts is neither discretely segmental in time, nor categorial in the sense of involving an inventory of discretely different elements. Segmentalised phonetic description is further challenged by instrumental evidence that neutralisation may be phonetically incomplete; and that segmental contrast may be cued over domains as large as the stress foot. Phonetics needs a more explicit statement of the relation of segments to articulation and to perception.

INTRODUCTION

Throughout the history of modern phonetics the phone-sized segment has been crucial. True, other elements such as syllables and features have periodically competed for attention; but the centrality of the phone is such that even these alternative elements are often talked of as combinations of, or decompositions of, phones.

The phone-sized segment remains at the heart of phonetic description despite decades of instrumental research into articulation and acoustics demonstrating beyond doubt that discrete phones do not exist in a straightforward way in the speech event, at least as superficially observable. An x-ray film of speech, or a dynamic palatographic record, shows gestures for different segments overlapping and blending. And from the earliest speech synthesis it has been known, for instance, that the perceptual cues to a consonant are distributed at least over the adjacent vowels.

But the survival of the phone is not hard to explain. It is the basis of our only extensive model of general phonetic description, as embodied for instance in the alphabet of the International Phonetic Association. This in turn reflects the fact that phone- or phoneme-sized units provide the most generally applicable and revealing descriptions of the phonologies of languages.

Thus the phonetic sciences have proceeded in a somewhat schizophrenic state of mind, knowing that phones aren't really there, but at the same time they have to be there. The hope is generally that at some stage the relationship between segmentalised descriptions and the continuum of speech performance will become clear and well specified.

This paper draws together a number of cases where it seems that the tension between the discretely segmental description and the observable speech event is high enough to make the resolution of their relationship a priority.

DISCRETENESS ON TWO AXES

The traditional phone-based model of phonetic description implies discreteness on two axes.

Firstly, the phone symbols from left to right in a transcription imply a temporal sequence of discrete phonetic events. The strongest interpretation of this, with for instance all acoustic cues to a segment ending simultaneously and abruptly at a boundary with a following segment, is clearly falsified even by casual observation of spectrograms. Perhaps the weakest interpretation is one which allows overlapping in the realisation of phones, but still expects their implied sequencing to be respected in that the realisation of phone n will not extend later in time than that of phone n+1 nor earlier than that of phone n-1 (see Fig.1). For instance, if in the utterances [ski] and [sku] the friction of the [s] contains from its onset cues to the velar, and to the tongue+lip configuration of the vowel, the implied sequencing has been respected. If however the velar or is not cued until late in the friction, whilst the vowel configuration is cued from the start, the implied sequencing has been violated (because cues to segment 3 begin before cues to segment 2). The issue of 'proper sequencing' has probably not been addressed in quite this form in existing work on coarticulation.

Secondly, the phone symbols imply selection of phonetic events from a paradigm of discrete phonetic units. Thus [d] is either present in an utterance or it is not, and if it is, it is whole-heartedly [d] and not something which vacillates between being a [d] and being a [b]. Speech performance is thus implied to be categorial.

CONNECTED SPEECH PROCESSES

It is reasonable to expect that phone-segmental phonetic description would be under greatest stress with fluent connected speech. Carefully produced citation forms yield maximal phonological contrast, and come nearest to exhibiting a simple relationship between segmental representations and the physical event. In connected speech the explicitness of the realisation of phonological contrasts may be reduced in a variety of ways, including assimilation and deletion. These various reductions in explicitness have been termed connected speech processes (e.g. Barry 1985).

Linguistic phonetics has, perforce, described connected speech processes (CSPs) in phone-segmental terms: a phone is deleted, or changes into another phone (which, in the case of assimilation, more closely matches an adjacent phone in one or more phonetic dimension). It is unclear, however, whether the categorial change implied by this type of description reflects the facts of speech performance (production, or perception) since, as pointed out above, a phone-segmental representation inherently implies discreteness sequentially and paradigmatically.

To find out if assimilation involves a discrete change in production we can compare forms eligible for assimilation with forms which underlyingly contain the segment potentially created by the assimilation. For instance, when a speaker assimilates the place of articulation of the final stop in road to the following velar in the road collapsed, is the utterance then phonetically identical in every respect to the realisation of the rogue collapsed?

This question has recently begun to be studied using electropalatography (EPG). For instance, Barry (1985) shows that where a word-final alveolar precedes a word-initial velar there are three possibilities (see Fig. 2). The EPG display may show complete alveolar closure (no assimilation); it may be identical to the display for a matched utterance with an underlying velar word-finally (complete assimilation); or, crucially, in many tokens it shows that no closure is completed across the alveolar ridge, but nevertheless the sides of the tongue make contact far forward along the sides of the palate in a 'residual' gesture towards the alveolar ridge (partial assimilation). The occurrence of these types is influenced, though not directly determined, by speech rate.

The existence of partially assimilated forms is supported in a similar experiment by Kerswill (1985). The gradual nature of assimilation in production is at variance with the paradigmatic discreteness of phone based representations. In principle, articulation could be categorial in that a speaker either made a gesture sufficient to create a given configuration of the vocal tract, or did not make it. Instead, articulation appears to be gradual - in that some gestures are allowed to be present, but inadequate (from the point of view of the phonetic target, and probably from that of perception - as discussed later). Note that it is not simply the case that a gesture is being curtailed by increased rate; Kerswill (1985) shows that a speaker can speak fast but with relatively few reductions when asked to speak 'carefully'.

NEUTRALISATION

It appears that it is not only connected speech processes which put phone-segmental descriptions under strain. Recent instrumental work has suggested that in many long-accepted instances of phonological neutralisation there is, contrary to the traditional view of neutralisation, some phonetic realisation of the underlying (morpho-) phonological contrast.

Thus it has been argued that the underlying word-final voicing contrast is reflected in small, but measurable phonetic differences in languages where it is normally considered to be neutralised on the surface, such as German (Mitleb 1981, Charles-Luce 1985), Russian (Chen 1970), Polish (Slowiaczek and Dinnsen 1985), and Catalan (Dinnsen and Charles-Luce 1984). The dimensions of the realisation include the duration of the vowel preceding the stop, the duration of the stop, and the way in which these durations are affected by the class of sound at the beginning of a following word. Such evidence is not uncontroversial (e.g. Iverson and Fourakis 1984), but may at least be seen as raising the possibility that neutralisation in these languages is incomplete, rather in the way that the assimilations discussed above may be partial.

To the extent that the residue of the underlying contrast is in the preceding vowel, sequential discreteness of segments is violated, rather in the way argued by Chomsky (e.g. 1964) to make linearity an unfeasible condition on the relationship between phonemic and phonetic representations. It has recently been suggested, however, that such violations may be more extensive than implied by Chomsky's discussion of adjacent segments. Scott (1984) claims that American English listeners can differentiate potentially neutralised pairs of the writer-rider type, and that they do so on the basis of 'cues other than preceding vowel duration or the acoustic properties of the flap'. These cues include overall durational properties of the words and global differences in phonetic quality - e.g. that rider is more 'open mouthed' in its articulation than writer. Kelly and Local (1986) suggest, too, that the spectral cues to /l/ versus /r/ in English extend over appreciably larger domains than usually considered - perhaps as extensive as the stress foot.

PRODUCTION, PERCEPTION, AND PHONETICS

Might it be the case that phenomena which hit the limits of segmental description are of no interest to phonetics because they are not perceptible, and therefore of no communicative value? On the contrary, their perceptual status forces consideration of one of the major ambiguities of phonetic analysis.

The ambiguity is whether a transcription is a record of what is said, or what is heard. As long as these coincide, the ambiguity is unobtrusive. But if, for instance, it were the case that German speakers reliably produced a measurable difference in Rad-Rat but neither native speakers nor phoneticians could perceive it, what would the correct phonetic transcription for the pair be?

The evidence as yet is inconclusive. Port and O'Dell (1985) report, for German, 59% correct identification of (incompletely) neutralised lexical items, compared with 50% as chance. Experiments are proceeding in Cambridge to test whether listeners are able to exploit perceptually the residual articulations of partial assimilations. And in a case of a phonological merger in progress, Costa and Mattingley (1981) show that subjects exibit a residual vowel duration difference in New England cod versus card, but are unable to exploit it perceptually.

On the whole it seems probable that at least some cases will emerge where reliable production differences realising phonological contrasts are not perceived. The following table sets out some of the logical possibilities. In the three columns + or - indicates whether or not a distinction is (A) articulatorily realised by a speaker, (N) perceived by a native speaker, and (P) perceived by a well-trained phonetician in 'analytic' mode.

	A	N	P
(a)	+	+	+
(b)	+	+	-
(c)	-	+	-
(d)	+	-	+
(e)	+	-	-

(a) represents the unproblematic ideal.
(b) the native speaker and listener are coping fine with the distinction; the phonetician must try harder.
(c) the native listener hears a distinction which isn't there; this makes sense within a view such as that of Chomsky and Halle (1968) where the phonetic percept is partially determined by higher level linguistic knowledge.
(d) native users produce a distinction without reliable perception, while it may be salient enough for phoneticians to identify; some 'merger in progress' cases appear to fit here.
(e) in other cases the measured effect may be too small for the phonetician.

Hitherto it has been convenient to regard a phonetic representation as a linguistic construct, independent of articulatory and perceptual domains, but with definable (if as yet undefined) and equivalent relationships to each. The emergence of evidence of a lack of congruence between what a speaker produces and what he perceives may force a reappraisal of precisely what a phonetic representation should account for.

CONCLUSION

Phonetic description has revolved around the phone-sized segment. This construct is essentially discrete both sequentially and paradigmatically.

Sequential discreteness has long been recognised not to characterise any aspect (acoustic, articulatory) of the speech signal. The questions which seem currently worth pursuing are how extensive the influence of a segment is in time; and, perhaps, as summarised in Fig. 1, whether even proper sequencing is preserved in the speech signal.

Do speakers behave as if segments represent categorial choices? Apparently not; in environments with the potential for place of articulation assimilation a gradation of assimilation occurs.

Categories may be a function of hearing, rather than speaking. The continuum of behaviour from no place assimilation through partial to complete assimilation may turn out to yield a categorial perceptual boundary somewhere in the 'partial' region. But on the other hand it is possible that no perceptual boundary will emerge because, as with the cod - card case, listeners can't exploit the acoustic details.

A consideration of the limits of segmental description, then, inevitably leads to consideration of the status of the categories which phone-segments imply, and of the representations which they comprise. If the disparity between production and perception which is hinted at by work cited here is confirmed, the general conception of phonetic analysis will have to be radically revised and its relation to aspects of speech performance made explicit.

REFERENCES

BARRY, M. 1985 A palatographic study of connected speech processes. Cambridge Papers in Phonetics & Exp'l Linguistics 4.
CHARLES-LUCE, J. 1985 Word-final devoicing in German: effects of phonetic and sentential contexts. J.Phon. 13, 309-24.
CHEN, M. 1970 Vowel length variation as a function of the consonantal environment. Phonetica 22, 129-59.
CHOMSKY, N. 1964 Current issues in linguistic theory. In: J.A. Fodor & J.J. Katz (eds), The Structure of Language.
COSTA, P.J. & MATTINGLY, I.G. 1981 Production and perception of phonetic contrast during phonetic change. Status Rep. on Sp. Res. SR-67/68, 191-6. Haskins Labs.
DINNSEN, D.A. & CHARLES-LUCE, J. 1984 Phonological neutralisation, phonetic implementation and individual differences. J.Phon. 12, 49-60.
FOURAKIS, M. & IVERSON, G.K. 1984 On the 'incomplete neutralization' of German final obstruents. Phonetica 41, 140-9.
KELLEY, J. & LOCAL, J.K. 1986 Long-domain resonance patterns in English. To appear in IEE Proceedings of Conference on Speech Input/Output, London.
KERSWILL, P.E. 1985 A sociophonetic study of CSPs in Cambridge English: an outline and some results. Cambridge Papers in Phonetics & Exp'l Linguistics 4.
PORT, R.F. & O'DELL, M.L. 1985 Neutralization of syllable-final voicing in German. J.Phon. 13, 455-71.
SLOWIACZEK, L.M. & DINNSEN, D.A. 1985 On the neutralizing status of Polish word-final devoicing. J.Phon. 13, 325-41.
SCOTT, D.R. 1984 More on the /t/:/d/ distinction in American alveolar flaps. JASA 75 (Suppl. 1), S66.

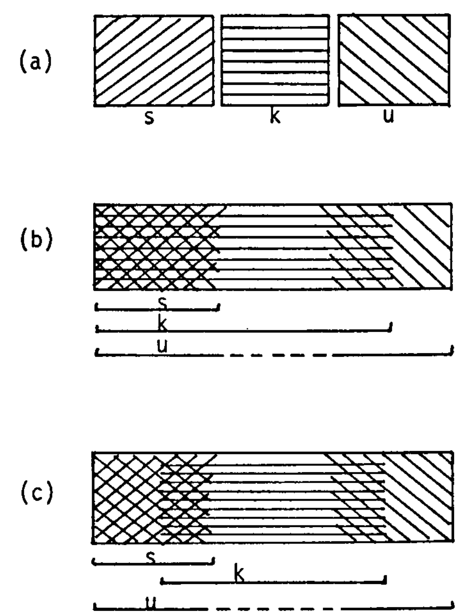


Fig 1 (a) discrete phones, as not found in the speech event; (b) implied sequence of phones in [sku] preserved; and (c) violated, since cues to the vowel precede those to the stop.

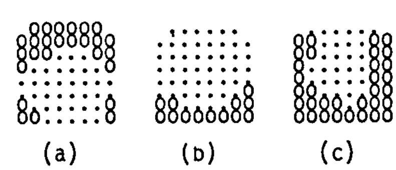


Fig 2 EPG displays taken from (a) unassimilated alveolar; (b) alveolar completely assimilated to following velar; (c) alveolar partially assimilated to following velar, showing maximum 'residual' alveolar gesture.

SOME ASPECTS OF 21 SPOKEN BULGARIAN CONSONANTS PERCEPTION

RADKA KURLOVA

Laboratory of Applied Linguistics
Institute for Foreign Students, 1111 Sofia, Bulgaria

ABSTRACT

The perceptual organization of 21 spoken Bulgarian consonants and the distinctive features have been determined using similarity and dissimilarity data drawn from two perceptual experiments with 200 Bulgarian native speakers.

INTRODUCTION

In the past decade research in speech perception has utilized information-transmissions, cognitive strain in short-term memory, linguistic, psychophysical, and reaction-time methods to gain insight into speech processing. In addition, there are many variants to the methods of measurement in psychophysics which include absolute judgement or direct estimation, scaling of paired comparisons, and triadic comparisons. An extensive review of many of these approaches can be found in Singh /22/ and Dauhauer and Singh/6/. In studies investigating the constituents of the phonemes the common elements have been the articulatory and acoustic features of the input stimuli. In general, the distinctive features have been consistently retrieved. Collectively, these studies appear to have established their psychological reality and perceptual independence relative to the input stimuli /29,28,17,21/.It has been determined as well that a hierarchical structure exists within the phonological domain of distinctive features/25/. Utilizing some aspects of the above mentioned methods and on the basis of experimental results /2,3,1,5,18,19,20,26,27/ a model of the phoneme as theoretical construct was developed/14/. The phoneme is represented as a three-space unity in which the physical reality of the speech unit, the phonological construct of the phoneme, and the perceptual speech sound space of Subjects are described as sets of acoustic, distinctive, and psychological features, respectively. It is suggested that relations and correspondences exist among all types of features and spaces.

The acoustic features could be represented by one or more physical properties of the speech unit (segment) which are changeable in time. Sources of these changes could be several physiological and geometrical parameters, as well as some physical phenomena (the form and size of the vocal tract, "basis of articulation", the transition from one target configuration to the next). All processes attending articulation and coarticulation contribute to the variation of the acoustic features on the time-axis too. Our understanding of the acoustic feature character is very close to Stevens's view /25/. We assume that:1.An auditory system could give a distinctive response not only to the sound itself, but also to each physical property of the sound and its change in time, according to the psychophysical laws. 2.There are many invariant acoustic properties(physical ones) associated with each acoustic feature. 3.The simultaneous appearance of some physical properties and their variations could cause changes in the perception of other physical properties (a high frequency signal of great intensity is perceived as a lower frequency signal). A support of the third assumption has been found out in an investigation of the Bulgarian vowels. For the acute vowels /i/ and /e/ the third formant F3 influences the first formant frequency F_1/14/. The acoustic features can be measured objecively. They form an n-dimension physical space with its axes corresponding to the number of features. The allophones corresponding to the phoneme variants can be presented as a set of points in a fixed region of that space. The distinctive features characterize the phoneme as linguistic construct/9/. Each distinctive feature has its acoustic and psychological correlates. The type of the acoustic correlate depends on the phoneme in which the distinctive feature is realized. Up to some limits, the variations of the acoustic features (correlates) cause changes in the grade of the distinctive feature in the phoneme. In other words, the quality and variations of the physical properties of the acoustic correlates are transformed by the phonological system/15/ into an estimation of the distinctive feature gradation. The degree of the distinctive features in phonemes which are found in different languages is different due to phonological system distinctions.

The psychological features are characteristics of the perceptual speech-sound space of the Subjects. These features have correlates both in physical and phonologic al spaces and are represented as orthogonal axes. Their number defines the space dimensionality ana depends on the properties significant of the phoneme classification. The "psychological phoneme" occupies well confined "psychological region"which is invariable with respect to the psychological axes even in case the coordinate system is rotated or translated. The phonemes of the vocal and consonant systems can be studied within the framework of the model. Relations among physical, distinctive and psychological features of Bulgarian vowels have been found out. It has been established /4/ that the perceptual space of the spoken vowels is two-dimensional and the second formant frequency F_2 is an acoustical correlate of the distinctive feature grave-acute and the first psychological axis. As a result of another investigation/12,13/ the first formant frequency F_1 has been defined as an acoustic correlate of the distinctive feature compact-diffuse and of the second psychological axis (r=-0,97). The analysis of experimental data /7,8/ reveals that the perceptual space for whispered and sung vowels is two-dimensional. All spoken, whispered, and sung allophones of each vowel perceived were situated in a fixed region. The results support the hypothesis of the psychological reality of the phoneme (the "psychological phoneme") and we have good grounds for considering the phoneme not only as "the highly serviceable tool by which we describe our speech samples" and "a basic pragmatic function in speech"/11,p.59/ but also as a cognitive structure with its own complex dimensions.Twenty-one Bulgarian consonants have been studied by using multidimensional scaling technique/5/. The purpose of the present study is to determine the organization of 21 spoken Bulgarian consonants and their features in the perceptual space.

METHOD

Subjects. Subjects were randomly selected students from four classes of the Technichal School of Electronics in Sofia(N==100) and from four classes(N=100) of the Polytechnical School in Roman. All the students were 16-18 years old Bulgarian native speakers.

Stimuli. A stimulus set consisted of 210 pairs of CV-syllables. In each syllable the Bulgarian consonants /b/,/v/,/g/, /d/,/ʒ/,/z/,/k/,/l/,/m/,/n/,/p/,/r/,/s/, /t/,/f/,/h/,/ts/,/dz/,/tʃ/,/dʒ/, and /ʃ/ were uttered with the vowel /ə/ in the way they were spelled in the Bulgarians alphabet. The pairs were recorded with a microphone feeding high-fidelity stereo tape recorder "Jupiter"-202. The stimuli were uttered at a confortable loudness level by a female Bulgarian native speaker born in N West Bulgaria.

Procedure. Experiment I. The experimental set was presented auditory to a hundred students from the Technical School of Electronics. The Subjects were asked to give similarity judgements for each consonant pair using a 7-point scale whose categories were marked verbally by phrases (and by corresponding numbers running from 1 to 7) in the direction of increasing similarity.The scale was anchored on the left with the phrase "Not at all similar"(number 1) and on the right with the phrase"Very similar" (number 7). The Subjects were instructed to write down on the answer sheet their assessments of similarity between the consonants in each pair using the respective numbers. Experiment II. The same experimental set was presented auditory to a hundred students from the Polytechnical School in the manner described above. Only the scale was anchored with the phrase "Not at all different on the left (number 1) and "Very different" on the right (number 7).

RESULTS

Two symmetric matrices (of similarity - Table 1a and of dissimilarity - Table 1b,respecively) were obtained as a result of the two experiments. The matrices were analyzed by the method of Johnson's Hierarchical clustering scheme. The hierarchical clustering tree in Fig.1 proceeds from the analysis of the similarity matrix. The branches of the tree differ in lenght as contrasted with the common hierarchical clustering tree.

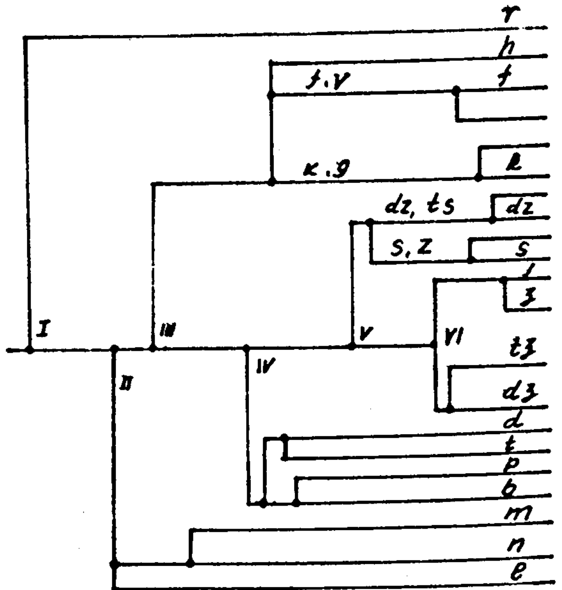


Fig.1. Representation of similarity between Bulgarian consonants as hierarchical clustering scheme

Table 1a. Matrix of Similarity for 21 Bulgarian Consonants
/Scores run from 1(lowest similarity) to 7(highest similarity)

1b \ 1a	b	p	v	f	g	k	d	t	ʒ	ʃ	z	s	dz	ts
b		5,25	3,96	3,53	3,55	2,99	4,49	3,92	2,57	2,55	2,51	1,78	2,89	2,61
p	3,27		3,59	4,20	3,26	3,02	3,48	3,71	2,17	3,31	2,09	2,07	2,24	2,81
v	4,75	4,70		5,72	2,79	2,26	3,12	2,56	2,26	2,76	2,46	2,51	2,47	2,27
f	4,39	4,26	3,11		3,21	2,85	2,91	3,65	2,67	4,14	2,76	3,63	3,01	3,56
g	4,48	4,65	4,87	4,47		5,89	3,48	3,38	2,86	1,95	2,64	2,49	3,17	2,63
k	5,29	4,91	5,47	5,14	3,39		3,11	3,18	2,15	2,73	2,22	2,65	2,02	2,36
d	4,59	4,58	4,78	4,81	4,42	5,33		5,17	2,80	2,24	3,42	2,57	4,36	2,66
t	4,29	4,26	5,14	4,85	5,76	5,08	3,34		2,98	3,69	2,40	3,40	3,34	4,30
ʒ	5,29	5,56	5,58	5,07	5,27	5,33	4,85	5,43		5,97	4,75	3,33	4,36	3,37
ʃ	5,23	4,64	5,15	4,17	5,26	5,09	5,00	4,06	3,12		3,74	5,03	3,33	3,76
z	5,10	5,22	5,45	5,35	5,11	5,26	4,60	5,43	3,45	4,29		5,76	5,45	4,53
s	5,63	5,31	5,24	4,77	5,26	5,28	5,06	4,87	4,73	3,60	3,02		4,95	5,32
dz	4,95	4,90	5,42	5,26	4,86	5,38	4,54	4,92	3,54	4,68	3,90	3,86		5,92
t	5,65	5,25	5,42	4,83	5,37	5,41	5,22	4,43	4,63	4,38	3,93	3,19	2,98	
d	4,82	5,36	5,47	4,85	4,90	5,58	4,48	4,93	3,54	4,99	3,80	4,91	3,47	4,09
t	5,42	5,27	5,48	4,97	5,14	4,94	5,01	4,79	4,57	3,57	4,67	4,80	4,07	3,96
h	5,30	4,41	4,65	3,52	4,71	4,27	5,05	4,03	5,17	3,79	4,92	4,60	5,16	4,92
m	4,99	4,88	4,91	5,75	5,43	5,18	5,32	5,45	5,83	5,02	5,61	5,28	5,51	5,91
n	4,86	4,64	4,99	5,04	5,07	5,04	4,91	4,85	5,46	5,05	5,52	5,35	5,33	5,61
l	4,75	5,06	5,04	5,16	5,26	5,20	5,39	5,25	5,48	5,20	5,38	5,63	5,67	5,69
r	5,72	5,57	5,35	5,15	5,57	5,60	5,93	5,61	5,68	5,23	5,51	5,54	5,72	5,57

Table 1b. Matrix of Dissimilarity for 21 Bulgarian Consonants
/Scores run from 1(lowest difference) to 7(highest difference)/

The 21 Bulgarian consonants are classified in six clusters /Fig.1) which are equivalent to different consonant classes. The hierarchical tree in Fig.2 is an upshot of the dissimilarity matrix analysis. The branches are represented by both the correlative pairs/dz,ts/,/s,z/,/ʃ,ʒ/,/f,v/ and /h/, /d,t/,/b,p/ and the groups/n,m/./l/./r/.

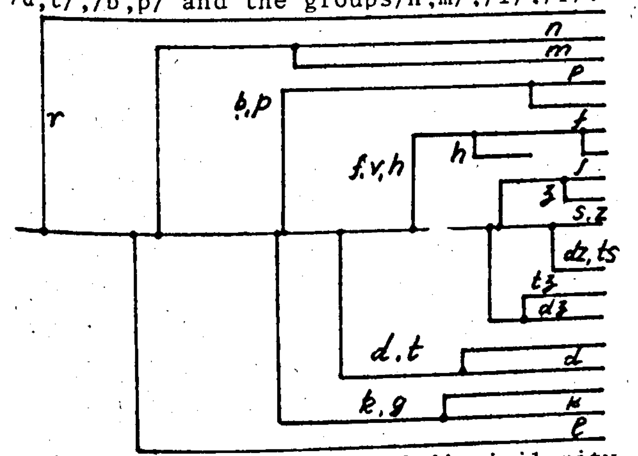


Fig.2. Representation of dissimilarity between Bulgarian consonants as hierarchical clustering scheme

DISCUSSION

In this section we relate the present results to the representation of consonants in terms of clusters and trees, and the concepts of family resemblance and prototypicality on the basis of Tversky's contrast model/23/. There is well-known correspondence between the classes and features or properties of the objects belonging to the class. This correspondence provides a direct link between the clustering approach to the representation of proximity data and the contrast model. The feature tree can be interpreted as a hierarchical clustering scheme

Table 2.Characteristics of Bulgarian consonant classes

Class	Family resemb.R	Prototypicality P
I	2,22	
II	3,44	2,38n,2,36m,2,13l
III	3,90	3,25f ,3,25h,3,12k,3,11g ,2,87v
IV	4,34	3,41b,3,28d,3,20t,3,11p
V	5,08	3,89ʃ ,3,87dʒ ,3,81ʒ ,3,66tʃ
VI	5,32	4,08dz,4,01s,4,94ts,4,93z

Note. The calculations of R and P were conducted according to /23/

Table 1 (continued)

d	t	h	m	n	l	r
2,78	2,50	2,70	2,90	2,70	2,61	1,77
1,95	2,66	3,58	3,08	2,87	2,67	2,20
2,20	2,28	3,23	2,72	2,84	2,91	2,23
2,67	3,14	4,81	1,85	2,43	2,80	2,59
3,12	2,85	3,64	2,31	2,94	2,20	2,14
2,26	3,05	4,58	2,51	2,94	2,57	2,17
3,79	2,72	2,81	2,31	3,17	2,19	2,05
3,46	3,84	4,10	2,67	2,48	2,29	2,50
5,33	3,93	3,09	2,20	2,16	1,90	2,01
4,53	5,07	4,42	2,35	2,54	2,38	2,33
4,73	3,01	2,96	2,04	2,37	2,16	2,09
3,35	4,95	3,82	2,28	2,22	2,30	2,23
5,32	4,58	2,91	2,02	2,58	1,70	1,87
4,61	4,55	3,57	1,88	2,41	1,71	2,32
	5,64	2,71	2,09	2,24	1,86	2,08
3,29		3,26	2,12	2,42	1,79	2,10
4,87	4,63		2,04	2,83	1,80	2,71
5,39	5,72	5,39		3,91	3,17	2,29
5,16	5,20	4,92	4,25		3,24	2,21
5,46	5,37	5,47	5,00	4,83		2,44
5,20	5,42	5,32	5,73	5,26	5,40	

where each arc lenght represents the weight of the cluster consisting of all objects that follow from that arc. It is known that similarity is a relation of proximity that holds between two objects or concepts, prototypicality (P) is a relation between an object (concept) and a class, family resemblance (R)is a network of similarity relations that link the various members of the class. Clusters form so as to maximize similarity of objects within the class and dissimilarity of objects from different classes, therefore the class with higher family resemblance separates earlier in clustering. Table 2 reflects the measures of family resemblance of the consonant classes, and the prototypicality of the class members. The relation between family resemblance and each cluster is represented graphically in Fig.1. The arc lenght of the clusters is inverse to R and shows that the class with the highest R forms first. The correlative pair including the class member with the highest prototypicality attracts the nonpaired members (the pair/f,v/ attracts /h/). The order of the correlative pairs separation from the tree stem (Fig.2) is closely related to the pair similarity and the difference (S-D) between similarity and dissimilarity. The correlation between pair similarities and differences (S-D) is -0,94. There is no correlation between pair similarities and dissimilarities (r =-0,50), and between pair dissimilarities and differences (S-D)(r =076).These findings imply that the salience of the feature changes in the pair so that difference(S-D)and similarity remain in linear relation.

The organization of the 21 Bulgarian consonants in the perceptual space can be well interpreted in terms of the proposed phoneme model. In support we would like only to mention that there are relations among psychological axes, family resemblance, features, and physical properties of the consonants, and that time is the link connecting difference (S-D), order of pairs separation, and the distinctive feature of voicing.

REFERENCES

/1/Бондарко Л.В.,"Слоговая структура речи и диф.пр.фонем",Авт.докт.дис.,Л.,1969.

/2/Chistovich L.A.,"Auditory Proc.of Sp" Proceedings of the 9-th ICPhS,vol.1,Cop.'79.

/3/Chistovich L.A,Ogorodnikova E.A."Temp. Proc. of Spectra in Vowel Perception",Speech Communication,v.1,No.1,1982.

/4/Gerganov E.et al"MDSCAL of 21 Bul.Consonants",2ndNat.Conf."Acoustics'75",1975.

/5/Gerganov E.et al."A Model for Percept. of Sp.Sounds,Nat.Conf."Acoustics'75",1975.

/6/Dauhauer J.and Singh S.,"MDSp.Percept. by the Hear.Impaired:A Treat. on DF,NY,1975.

/7/Gerganov E.and Kurlova R.,"Perc.of Bul. Vowels Sung with Diff.Fo"Yearbook of Inst. for Foreign Stud.,v.1, 1983.

/8/Gerganov E.and Kurlova R.,MDSCAL of Bul.Vowels,Yearbook of IFS,v.2, 1983. „

/9/Jakobson R.,Fant G.,and Halle M.,Prel. of Sp.Analysis"Cambr.,MIT Press,1963.

/10/Johnson S.C."Hierarchical Clustering Schemes",Psychometrica,32,1967,241-254.

/11/Kent R.D."The Segm.Organ.of Sp."in: The Prod.of Sp.(ed.byP.F.MacNeilage),NY,'83.

/12/Kurlova R."Bul.Vowel Recogn."Intern. Conf."Robcon-2",Varna,1983.

/13/Kurlova R.,"Phys.Space of Bul.Vowels", Yearbook of IFS,v.3,1985.

/14/Кърлова Р.,"Експер.фонет.и перцепт. изсл.на бълг.гл.",Авт.кандид.дис.,С.,1985.

/15/Kurlova R.,"Transf.of Bul.vowels", Yearbook of IFS, v.4,1986.

/16/Kurlova R.and Gerganov E.,"Appl.of Meth.of Princ.Comp.and HCl.Scheme,12th Europ. Meet.of Statisticians,Varna,1979.

/17/La Riviere C.et al.,"The Concept.Reality of Select.DF,JSHR,17,122-133.

/18/McNeil D.,and Linding K.,"The Percep. Real.of Phon.,Syll.,Words and Sent.,JVLVB, v.12,No.4,1973.

/19/Randy D.et al."Perceiv.V.in Isol.and Conson.Context,JASA,v.69,239-248,1981.

/20/Singh S.and Woods D.,"Precept.Str.of 12 Am.Engl.Vowels,JASA,v.49,1981.

/21/Singh S.et al.,"Percept.Str.of 22Pre-Voc.Engl.Consonants",JASA,v.52,1972.

/22/Singh S.,"Dist.F.:A Meas.of Conson. Percept.",Univ.Park Press,Baltimore,1975.

/23/Tversky A.,"Features of Similarity", Psychol.Review,v,84,No.4,1977.

/24/Wickelgren W.,DF and Err.in STM for Engl.Cons.,JASA,v.39,1966.

ELECTRO-PALATOGRAPHIC STUDIES ON JAPANESE INTERVOCALIC /r/ AND /d/

MASAYUKI SAWASHIMA AND SHIGERU KIRITANI

Research Institute of Logopedics and Phoniatrics,
Faculty of Medicine, University of Tokyo,
Tokyo, Japan

ABSTRACT

Electro-palatographic studies on the tongue palate contact patterns have been conducted on Japanese /d/ and /r/ in /VCV/ sequences in a carrier sentence. Subjects were 3 adults of Tokyo dialect. Complete stop closure at the anterior palate was seen for /d/ while many /r/ samples showed incomplete closure. Some of /r/ samples showed the anterior contact separate from the contact at the lateral part of the palate. Time curves of the anterior contact for /r/ revealed smaller area and shorter time span than those for /d/.

INTRODUCTION

Japanese /d/ is a stop consonant with formation of closure at the back of teeth and/or the alveolar ridge, while it is generally said that Japanese /r/ in intervocalic position is realized as a tap or a flap, with the tip of the tongue making one tap against the alveolar ridge. This stop-flap opposition implies that the palato-lingual contact is shorter in duration and also smaller in area for /r/ than for /d/.

In the field of experimental phonetics, use of the electro-palatography is considered to be one of the most powerful approaches for elucidating articulatory characteristics of the two sounds. Electro-palatographic finding of shorter duration and smaller area in articulatory contact has already been reported[1]. However, the data were quite limited and a more systematic study was needed.

In the present paper, results of our electro-palatographic study of tongue-palate contact patterns of Japanese intervocalic /d/ and /r/ in varying vowel contexts are presented.

EXPERIMENTAL PROCEDURES

Three native Japanese speakers of the Tokyo dialect served as subjects. None of the subjects reported any speaking disabilities. Test words were meaningless sequences of the form $/V_1CV_2V_1CV_2/$ (V_1=i, e, a, o, u; V_1=e, a, o; C=d, r). The test words were embedded in the carrier sentence /Sorewa ____ desu/ (It is ____). Each of the test sentences was repeated ten times, with a flat accent for the test word, at a comfortable speaking rate for the subject. Thus, 20 utterance samples were recorded for a given $/V_1CV_2/$ sequence.

The artificial palates used in this study had 63 electrodes. Contact signals from the electrodes in the artificial palate were stored in a computer connected to a portable electro-palatograph unit at a rate of 64 frames/sec. When the subject read a test sentence and pushed the control button after each utterance, the data for a duration of one-second were stored in the computer. The speech signals were also sampled by the computer at a rate of 64 frames/sec after rectification and integration over a 16 msec time window. The stored data were reproduced and observed in slow motion on an oscilloscope. The plotting of the necessary contact patterns was printed out by a high-speed line printer.

RESULTS AND COMMENTS

1. Maximum contact patterns

For each of the utterance samples, successive palatographic frames indicating the time course of the articulatory tongue-palate contact for the pertinent consonant were obtained. The peak articulatory contact was identified as the frame showing the maximum contact (maximum contact pattern) in the frame series. Maximum contact patterns were collected for all the utterance samples. With these maximum contact patterns, we constructed a contact pattern which consisted of the electrodes showing contact in more than 10 (50%) of the 20 repetitions, for each test word of each subject. This pattern was considered to be the average contact pattern for each test word in a given subject. The results are shown in Fig. 1.

In the figure, the patterns for /d/ and /r/ in the same vowel context are superimposed on the scheme of the artificial palate. The area demarcated by the thick line indicates the contact area for /d/, while the shaded area indicates that for /r/.

The average patterns reveal that, for /d/ there is a complete stop closure at the anterior margin of the palate for all of the vowel contexts in all three subjects. Also, there is little variation in the contact pattern among the different vowel contexts at the anterior part of the palate within each subject, while there is some context-dependent variability at the posterior part.

The average patterns for /r/ generally show a smaller contact areas than those for /d/. At the anterior part of the palate, there are many /r/ patterns which do not show complete closure.

Table 1 summarizes the contact area as defined by the number of on-electrodes in the maximum pattern for selected /d/-/r/ pairs of the test words. It is seen that for all the subjects /d/ shows a greater contact area than /r/.

Frequency of the occurrence of complete closure for 20 tokens of selected /d/-/r/ pairs of the test words are summarized in Table 2. It is noted that for all the subjects, most or more than half of the /d/ patterns show complete closure while more than half of the /r/ patterns do not.

The characteristic feature for Subj. 1 is that the contact at the anterior part shifts backward for /r/ as compared to /d/. This appears to occur

Utterance / Subject	ada/ara	ede/ere	odo/oro	Average
Subj. 1	25	28	22	25
	16	20	17	18
Subj. 2	22	28	21	24
	17	20	16	18
Subj. 3	24	27	26	26
	17	22	22	20
Average	24	28	23	25
	17	21	18	19

Table 1: The average number of on-electrodes in the maximum contact patterns for 20 tokens of selected test words.

Utterance / Subject	ada/ara	ede/ere	odo/oro	Average
Subj. 1	20	20	20	20
	1	6	7	5
Subj. 2	13	13	13	13
	6	1	4	4
Subj. 3	19	14	20	18
	2	10	14	9
Average	17	16	18	17
	3	6	8	6

Table 2: The frequency of occurrence of complete closure for 20 tokens of selected dest words.

only in the context of the back vowels for Subj. 2, while no such shift in the place of contact is observable for Subj. 3. Thus, there is some individual variation in the contact patterns for /r/.

Another feature of /r/ is that some of the /r/-patterns show contact at the anterior part separate from the contact at the lateral part of the palate. Whether this represents a specific tongue gesture for /r/ or not is an open question at this moment.

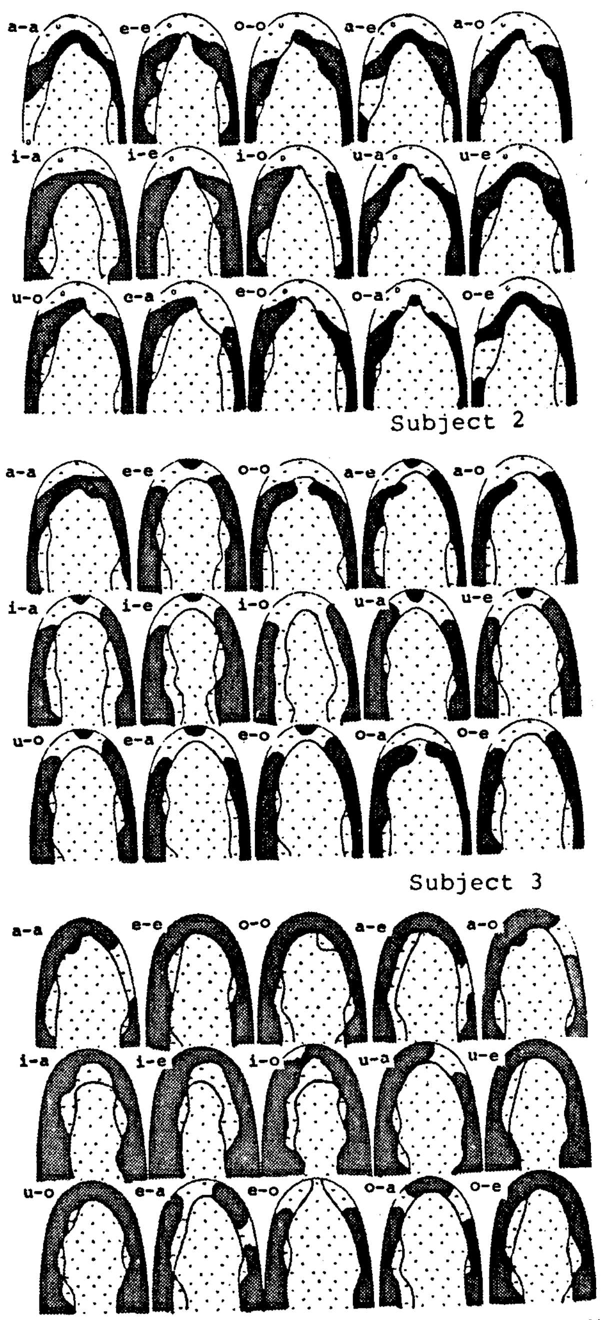


Fig. 1: Maximum contact pattern (averaged) for /d/ and /r/. The area demarcated by solid line is for /d/, the shaded area for /r/.

2. Time course of the tongue-palate contact

As described above, there was a complete stop closure at the anterior part of the palate for most of the /d/ patterns. The duration of the complete closure ranged from 2 to 4 frames out of 64 frames/sec. Some of the /r/ patterns also showed this stop closure. The duration of the closure in these cases ranged from 1 to 2 frames. This indicates that there is a difference in the time pattern, as well as the spatial pattern, of the tongue-palate contact between /d/ and /r/.

We then determined the average number of on-electrodes at the anterior part of the palate for 20 repetitions along the time course of each $/V_1CV_2/$ sequence, as shown in Fig. 2. In the figure, the ordinate of each graph indicates the number of on-electrodes and the abscissa the time axis. The time curve is demarcated by each frame of the palatogram, and the vertical line on the curve indicates the standard deviation. The dashed line indicates the contact for /d/ and the solid line that for /r/.

It should be noted that the area of the contact, i.e., the number of on-electrodes, is larger for /d/ than for /r/ throughout the time curse for all of the subjects and for all of the test samples. Also, it is apparent that /d/ shows a longer time span than /r/ both in peak contact and in the transition of the contact area. Thus, the /d/ and /r/ curves of Subj. 1 are clearly separated

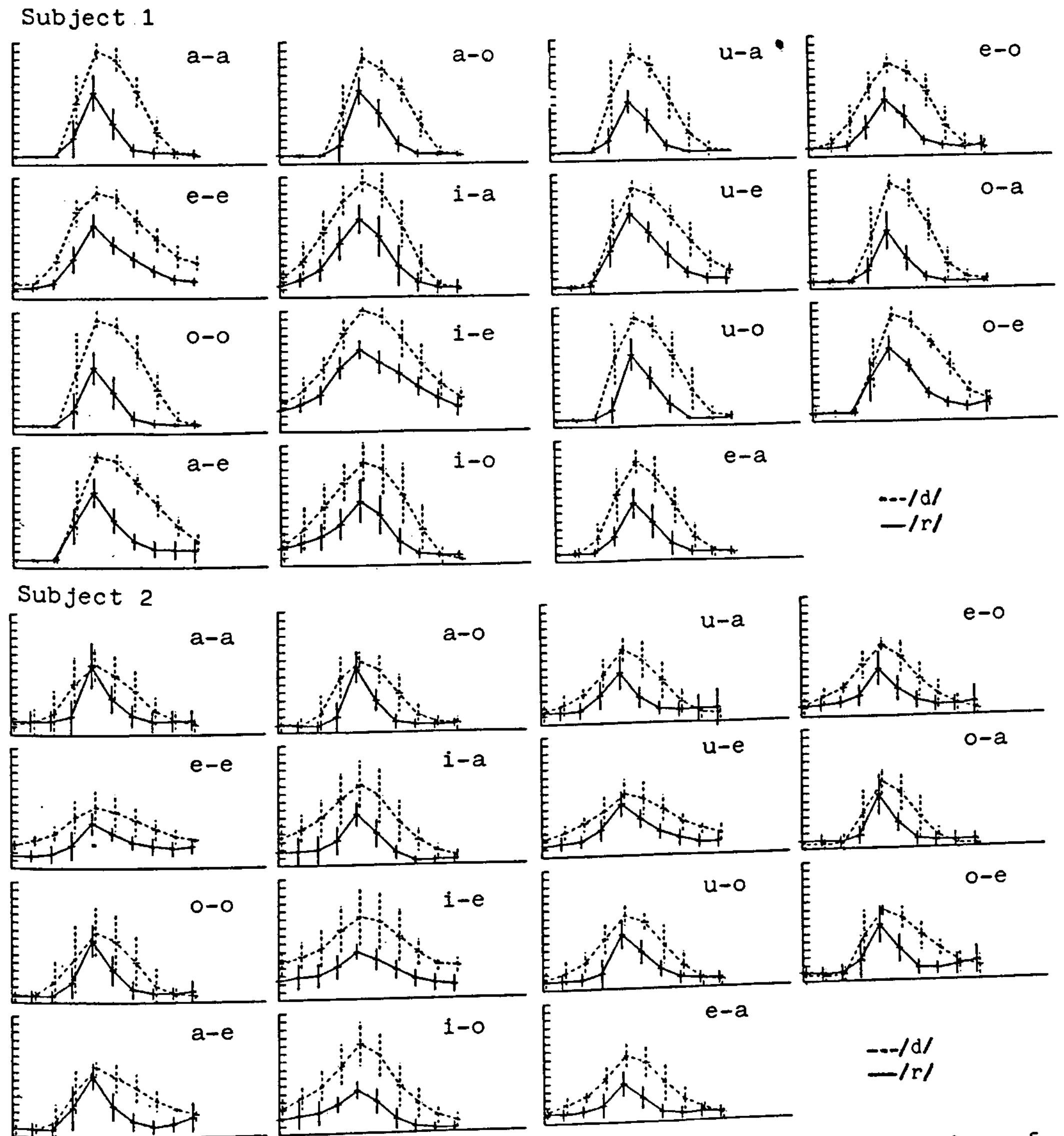


Fig. 2-1: Time curve of the area of contact as defined by the number of on-electrodes at the anterior palate for /d/ and /r/ for Subj.1 and Subj.2. Dashed line is for /d/, solid line for /r/.

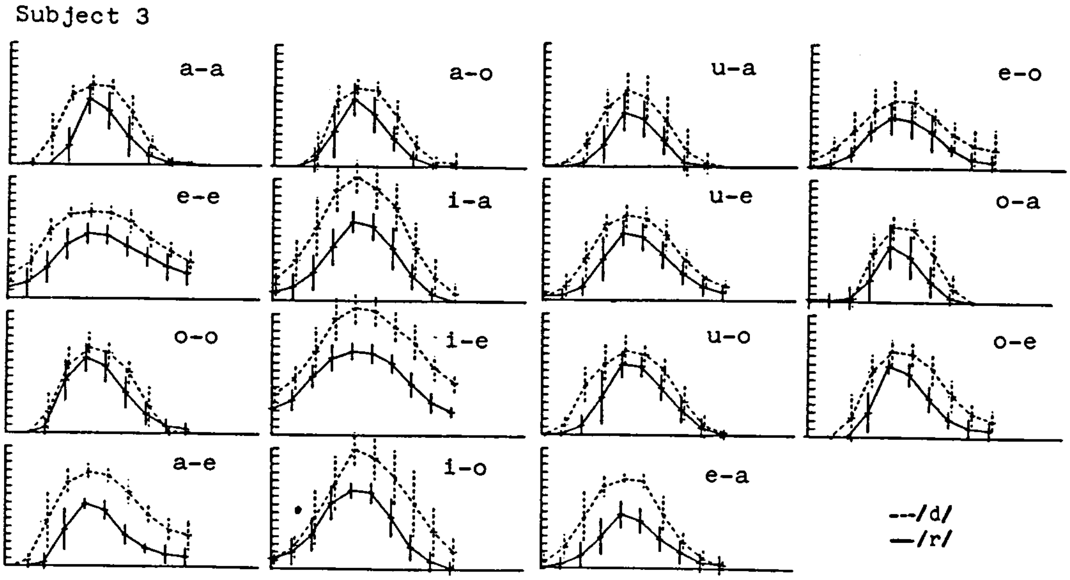


Fig. 2-2: Time curve of the area of contact as defined by the number of on-electrodes at the anterior palate for /d/ and /r/ for Subj.3. Dashed line for /d/, solid line for /r/.

from each other for all of the vowel contexts. Some of the curves of Subj. 2 show that the peak values of contact for /d/ and /r/ are comparable to each other. In these cases, however, the /r/ curves show a much steeper slope before and after the peaks than the /d/ curves, indicating a faster transition to and from the peak contact for /r/. The time curves of /r/ for Subj. 3 present rather similar contours to those for /d/, the contacts for /r/ showing smaller values than those for /d/. Thus, the distinction between /d/ and /r/ appears to be less evident in Subj. 3 than in Subjs. 1 and 2, as far as the tongue-palate contact pattern is concerned.

SUMMARIES

Electro-palatographic study was conducted on Japanese intervocalic /d/ and /r/. The results were summarized as follows:

1) Maximum contact pattern revealed that the contact area was greater for /d/ than for /r/. It was also noted that most of the /d/ patterns showed complete closure at the anterior palate while many of the /r/ patterns did not.
2) Some of the /r/ patterns showed the anterior contact separate from the lateral contact along the teeth ridge, which was never seen in the /d/ patterns.
3) Time course of the anterior contact revealed a shorter time span of articulatory contact for /r/ than for /d/.
4) There appeared to be greater individual variation in the articulatory contact for /r/ than for /d/, which resulted in some individual variation in the difference between /d/ and /r/.

REFERENCES

[1] Fujimura, O., Tatsumi, I. F. and Kagaya, R.: Computational processing of palatographic patterns. J. Phonetics, 1; 47-54, 1973.

ACKNOWLEDGEMENT

This study was supported in part by a Grant-in-Aid for Scientific Research (No. 59101003) from the Japanese Ministry of Education, Science and Culture.

OBJEKTIVE BEWERTUNG VON /S/-ALLOPHONEN

EBERHARD STOCK UWE HOLLMACH

Bereich Sprechwissenschaft
Martin-Luther-Universität
Halle, DDR 4020

ZUSAMMENFASSUNG

Es wird über die Entwicklung eines in der Praxis verwendbaren, möglichst billigen und möglichst einfachen computergestützten Verfahrens zur Rationalisierung sprechwissenschaftlich-phoniatrischer Routineuntersuchungen bei Studienbewerbern berichtet. Die Aufstellung der Vergleichsnorm und die Durchführung des Vergleichs werden diskutiert.

PROBLEMSTELLUNG

In der DDR bewerben sich in jedem Jahr zehntausende von Abiturienten an den Universitäten und Hochschulen mit dem Ziel, eine Ausbildung für einen sprechintensiven Beruf (z.B. als Lehrer, als Schauspieler, als Kindergärtnerin) zu absolvieren. Alle diese Bewerber müssen sich einer Tauglichkeitsprüfung unterziehen, die durch eine "Gemeinsame Anweisung der Miniterien für Volksbildung und Gesundheitswesen zur Beurteilung der Tauglichkeit für Berufe mit besonderer Stimm- und Sprechbelastung" aus dem Jahre 1974 geregelt wird. Nach dieser Anweisung sind folgende Untersuchungen durchzuführen:
(1) Sorgfältige Anamneseerhebung zur Einschätzung der Stimm- und Sprechleistung;
(2) Erhebung des HNO-fachärztlichen Status, gegebenenfalls mit audiologischen und röntgenologischen Überprüfungen;
(3) Ermittlung des Stimmstatus, d.h. des Stimmklanges, des Stimmeinsatzes, der Sprechstimmlage und der Steigerungsfähigkeit, gegebenenfalls mit Hilfe der Stroboskopie und der Pneumographie;
(4) Ermittlung des Sprechstatus, d.h. der Artikulation und des Sprechablaufs.

Als untauglich müssen Bewerber unter anderem dann eingestuft werden, wenn sie Stimm- und Sprechstörungen haben, die einer Behandlung nicht oder nur schwer zugänglich sind.

Diese Routineuntersuchungen sind außerordentlich arbeitsintensiv und zeitaufwendig, sie müssen dringend rationalisiert werden, damit die Behandlungskapazität der entsprechenden Einrichtungen nicht unzumutbar eingeschränkt wird. Unsere Arbeiten zielen darauf ab, für die Ermittlung des Stimm- und Sprechstatus ein möglichst einfaches und möglichst billiges computergestütztes Verfahren zu entwikkeln, das es ermöglicht, einen Teil der Untersuchungen Hilfskräften zu überantworten, ohne daß die Qualität der Beurteilungen dadurch eingeschränkt wird.

Wir haben aus zwei Gründen mit Arbeiten zur objektiven Bewertung von Sigmatismen begonnen: 1. Sigmatismen machen nicht nur im Kindesalter den größten Teil aller Lautbildungsstörungen aus, sondern sie sind auch bei Erwachsenen, wenigstens im deutschen Sprachbereich, als nahezu einzige funktionelle Stammelfehler weit verbreitet und werden von Kindern leicht imitiert. Studienbewerber mit auffälligem Sigmatismus sind deshalb für ein Lehrerstudium nicht tauglich.
2. Die subjektiv-auditive Bewertung der /S/-Realisationen ist vom Hörvermögen abhängig, das bekanntlich mit wachsendem Alter in der Höhe abnimmt, wodurch die Diskriminationsfähigkeit für korrekte bzw. unkorrekte /S/-Allophone eingeschränkt wird. Da die Zahl der Sigmatiker unter Lehrer- und Schauspielstudenten trotz Tauglichkeitsuntersuchung sehr groß ist und möglicherweise anwächst (in manchen Statistiken der letzten zehn Jahre werden bis zu 35 % der Studenten eines Matrikels als Sigmatiker ausgewiesen), sind wir der Beeinträchtigung des Hörvermögens und der Diskriminationsfähigkeit gesondert nachgegangen. Mit einem von U. Hollmach entwickelten Hochfrequenzaudiometer wurden in drei verschiedenen Altersgruppen Hörschwellenuntersuchungen durchgeführt. Es handelte sich (1) um zwölf Lehrerstudenten mit einem Durchschnittsalter von 21,2, (2) um zwölf Sprechwissenschaftler mit einem Durchschnittsalter von 31.7 und (3) um acht Lehrkräfte für Musik, alle über 5o Jahre alt mit einem Durchschnittsalter von 53.9. Das Audiometer kann in der Frequenz stufenlos zwischen 2o Hz und 22 kHz eingestellt werden, die Intensität

ist in 1-dB-Stufen regelbar. Die Umwertung der dB-Angaben in Phonzahlen und die Korrektur, die wegen der Eigenfrequenz des verwendeten orthodynamischen Kopfhörers und des für die Prüfung eingesetzten künstlichen Ohr erforderlich war, wurden mit einem speziellen Programm durch den Kleincomputer KC 85/3 des VEB Mikroelektronik "Wilhelm Pieck" Mühlhausen vorgenommen. Für die Prüfung wurden 15 Frequenzen ausgewählt. Die Mittel der Schwellenintensitäten ($\bar{x}$) und die jeweilige Standardabweichung (s) sehen abgerundet wie folgt aus:

kHz	Gruppe 1		Gruppe 2		Gruppe 3	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
o.4	6	4	7.8	4	3.8	6
o.7	8	4	6.4	7	3.4	3
1	4	4	4	6	1.5	3
1.2	o.9	3	2.4	5	2.1	4
1.5	4	7	6.2	5	4.9	6
2	5.9	4	4.9	6	12	8
2.5	6.7	5	7	8	13.8	11
3	6.3	6	9	9	16.6	8
4	5.6	9	8.8	7	19.3	12
6.3	9.6	6	12.7	9	27.9	23
7.6	5	6	1o.9	1o	38	66
9.5	6.6	6	14.8	8	48	63
1o.8	9.2	5	12.3	14	51	64
13	2	9	1o.2	12	1o1	82
16	16	11	48.8	81	–	–

Verständlicherweise wurde für die Gruppe der Studenten beim Mittel und bei den Extremwerten das beste Hörvermögen festgestellt. Bei der folgenden S-Bewertung urteilte sie am homogensten. Die zweite Gruppe zeigt die zweitbeste Mittelwertkurve; die Extremwerte liegen weiter vom Mittelwert entfernt, als in der ersten Gruppe. Die stärksten Ausfälle, besonders bei hohen Frequenzen, finden sich in der letzten Gruppe, in der aber eine Besonderheit zu beobachten ist. Bis 1,2 kHz haben die älteren Probanden eine um durchschnittlich 4 dB niedrigere Hörschwelle, als die beiden anderen Gruppen. Außerdem ist die hier nicht zusätzlich ausgewiesene Varianz in diesem Bereich für die Gruppe 3 deutlich geringer als bei den anderen Gruppen. Besonders bei den Studenten ist die Varianz bei 1,5 und 4 kHz sehr groß. Eine Erklärung hierfür steht noch aus.

Der Hörschwellenaudiometrie schloß sich sofort ohne Veränderung der Stellung des Kopfhörers ein Diskriminationstest mit Satzpaaren an, in dem gegenübergestellte /S/-Allophone auf Geräuschschärfe, auf Geräuschfarbe, auf Ähnlichkeit der Geräusche und auf Korrektheit der /S/-Realisation zu beurteilen waren. Um das Verhältnis zwischen den diskreten, von einem computergestützten Analysator ausgegebenen Spektraldaten der einzelnen /S/-Allophone und den ermittelten Hörschwellenwerten bestimmen zu können, wurde für jeden Probanden die Differenz zwischen Hörschwellenwerten und Spektraldaten ermittelt. Die entstehenden Differenzkurven ermöglichen Aussagen über die Wahrnehmungswahrscheinlichkeit für die einzelnen /S/-Allophone. Je positiver die Kurvenwerte sind, desto größer ist die wahrgenommene Lautstärke für die einzelnen Frequenzen und desto höher die Wahrscheinlichkeit, daß die wahrgenommenen Frequenzbereiche signalgerecht identifiziert werden. Erwartungsgemäß ergab sich, daß die Gruppe der Studenten die Sigmatismen am ehesten erkannte und am besten unterschied, daß die mittlere Altersgruppe auch eine Reihe von stumpfen S-Geräuschen als scharf beurteilte, daß diese sowie die letzte Gruppe den Sigmatismus stridens nicht erkannte und daß die über 5o Jahre alten Hörer die sigmatischen Allophone nicht sicher unterscheiden konnten. Diese Aussage bezieht sich selbstverständlich nur auf die auditive Beurteilung. In der Diagnose der sprechwissenschaftlichen oder phoniatrischen Praxis werden natürlich auch visuell aufgenommene Informationen genutzt. Trotzdem fordern diese Untersuchungsergebnisse, daß die erfahrenen älteren Kollegen, die in der Regel die für die Tauglichkeitsuntersuchungen zuständigen Einrichtungen leiten und dort aktiv tätig sind, nicht nur von Arbeitsaufwand entlastet, sondern auch hinsichtlich der Beurteilung der /S/-Realisationen unterstützt werden müssen. Nach unserer Vorstellung kann das auf sehr ökonomische Weise mit einem entsprechend programmierten Kleincomputer geschehen, der mit einem Analog-Digital-Umwandler bestückt ist und bei der Artikulation von Testwörtern oder kurzen Sätzen eine objektive Bewertung der /S/-Allophone, aber z.B. auch der Vokalartikulation im Fremdsprachenunterricht vornimmt. Der benutzte Kleincomputer KC 85/3 wird von uns vor allem deshalb für den Einsatz in der sprechwissenschaftlich-phoniatrischen Praxis empfohlen, weil er vielfältig nutzbar ist und z.B. durch RAM-Erweiterung und Zusatzmodul mit Textverarbeitungsprogramm auch die Rationalisierung der Anamnese und Diagnose gestattet.

AUFBAU DES SPEKTRALANALYSESYSTEMS

Für die Forschungsarbeiten mußte zunächst ein Spektralanalysesystem entwickelt werden, das es möglich macht, zusammenhängende natürliche Äußerungen mit einer bestimmten Mindestdauer, also "fließende Lautsprache", zu untersuchen. Dieses System besteht aus 4 Komponenten: einer Filterbank, einem Analog-Digital-Umwandler, einem Kleincomputer und der erforderlichen Software.

Mit der von U. Hollmach aufgebauten Filterbank wird die gesprochene Sprache in frequenzdiskrete Signale zerlegt, wobei der Abstand der einzelnen Frequenzkomponenten eine Viertel Oktave beträgt. Die Filterbank umfaßt 32 Kanäle, die parallel geschaltet sind und einen Frequenzgang von 8o Hz bis 18 kHz haben. Diese 32 Kanäle werden durch einen Analog-Multiplexer (AMUX) nacheinander durchgeschaltet, der das gleichgerichtete Signal seriell zum Ausgang führt. Die Integrationszeit bzw. das Zeitfenster ist zweistufig ausgelegt; Signale, die sich schnell verändern (z.B. Sprache), können mit einem Zeitfenster von 6 ms erfaßt werden, für quasi-konstante Signale (z.B. Stimmklang) steht ein Zeitfenster von 138 ms zur Verfügung. Während der AMUX an einem beliebigen Filterkanal das Analogsignal abfragt, würde an allen anderen Filterkanälen das Signal zeitlich versetzt weiterhin integriert werden. Das hätte eine Verfälschung der Signale zur Folge. Durch den Einsatz eines Analogspeichers wird die Ungenauigkeit verhindert. Im Analogspeicher steht der integrierte Filterausgangswert bis zur Abfrage zur Verfügung. In der Zwischenzeit kann an jedem Filter der zeitlich äquivalente Wert abgerufen werden.

Der Ausgang der Filterbank führt auf den Analog-Digital-Umwandler (ADU), mit dessen Hilfe die analogen Gleichspannungssignale in digitale, für den Computer verarbeitbare Signale umgesetzt werden. Der von uns benutzte ADU ist ein Zusatzmodul (M o1o) für den KC 85/3. Er besitzt eine Auflösung von 1o bit, das entspricht 1o24 Stufen.

Der KC 85/3 ist ein 8-bit-Rechner; er hat Vollgraphik und eine 16-farbige Displaystrukturierung. Ein besonderer Vorteil dieses Computers sind seine Erweiterungsmöglichkeiten. So kann selbstverständlich ein Drucker angeschlossen werden, und der RAM-Speicher kann bis zu 4 Mbyte aufgestockt werden.

Die Software wurde zur Hälfte in BASIC und für schnelle Abläufe in Maschinensprache (U 88o) geschrieben. Das Programm ist ein Grundprogramm für die Spektralisierung und Auswertung gesprochener Sprache und kann für jede spezielle Anwendung leicht erweitert oder umgestellt werden. Mittels Menütechnik können die graphischen Darstellungs- und Auswertungsvarianten benutzerfreundlich aufgelistet werden. Das Programm sieht eine farbige sonagraphische Darstellung mit 12 Farbwerten für die unterschiedlichen Intensitätswerte vor. Im Gegensatz zu den klassischen Sonagrammen mit ihrer Grauwertdarstellung ist die farbige Repräsentation übersichtlicher und genauer. Aus dem als Sonagramm erscheinenden Abschnitt der gesprochenen Sprache läßt sich insbesondere das stimmlose /S/-Allophon leicht herausfinden. Die optische Segmentierung erfolgt mit zwei Leuchtbalken. Aus den eingelesenen Daten kann sowohl ein Frequenz-Intensitätsdiagramm als auch ein Intensitätszeitdiagramm aufgebaut werden. Für die Beurteilung der /S/-Allophone ist das Frequenz-Intensitätsdiagramm vorteilhaft.

REFERENZMUSTER UND VERGLEICH

Die objektive Bewertung beliebiger idiolektaler /S/-Allophone geschieht durch den Vergleich mit einem Referenzmuster, das der Norm des Ausspracbestandards entspricht. Als Norm wird hier die von den Sprachbenutzern als dialektfrei und korrekt beurteilte Realisierungsvariante verstanden. Die Herstellung des Referenzmusters und der Vergleich orientieren sich an den Prozessen der Spracherkennung /1/. Die Feststellung und Beschreibung der Norm ist problematisch, so daß spezielle und noch nicht abgeschlossene Untersuchungen angesetzt werden mußten. Mit diesen Untersuchungen wurde in Kommunikationsexperimenten die Wirkung der einzelnen /S/-Allophone bzw. Sigmatismen als subjektive Bewertung durch den Hörer erfaßt. Methodisches Instrument waren standardisierte Polaritätsprofile, mit denen üblicherweise soziale Bezüge charakterisiert werden können. Die bisherigen Ergebnisse besagen, daß bei Kommunikation über Mikrofon (z.B. Rundfunk und Fernsehen) höhere Anforderungen gestellt werden und ein zu scharfes oder zu stumpfes S-Geräusch als auffällig und störend abgelehnt wird. In der face-to-face-Kommunikation und vor allem im Alltagsgespräch wird dagegen ein Sigmatismus stridens und selbst ein schwacher Sigmatismus addentalis sehr viel eher toleriert; hier ist die Norm also weniger streng. Je nach dem Berufsziel des Studienbewerbers wird man deshalb möglicherweise mit zwei Referenzmustern arbeiten müssen, einem Muster für die strenge Norm bei Mikrofonsprechern und einem zweiten Muster für die liberalisierte Norm bei Lehrern und Schauspielern. Das Referenzmuster entsteht, indem /S/-Allophone von mehreren Sprechern spektralisiert und die Spektraldaten gemittelt werden. Die einzelnen Allophone entstammen unterschiedlichen Phonemverbindungen in unterschiedlichen Äußerungen; jeder Sprecher wird durch eine Expertengruppe auditiv und visuell auf die Korrektheit seiner /S/-Realisation überprüft.

Bei der Diagnose bzw. der Erhebung des Sprechstatus werden /S/-Allophone aus Testsätzen mit dem gespeicherten Referenzmuster nach der Minimum-Distnaz-Strategie /2/ verglichen. Dieses statistische Verfahren nutzt die Beziehung zwischen dem Mittelwert der korrekten /S/-Realisationen und der durch die idiolektale Variation der Normsprecher entstehenden Streuung der Spektraldaten, wobei der Abstand zwischen dem jeweiligen Mittelwert und dem Streuungswert als die zulässige minimale Distanz bewertet wird. Überschreiten die Spektraldaten der /S/-Realisation eines zu diagnostizierenden Sprechers diese minimale Distanz, so wird diese /S/-Realisation zurückgewiesen. In unserer Forschung muß also praktisch die mögliche Unschärfe der /S/-Realisation bestimmt werden, die noch als korrekte Artikulation bewertet wird. Dabei ist ein Maß für die Ähnlichkeit aus den Distanzzahlen zu ermitteln.

Es besteht auch die Möglichkeit, die Wörter bzw. /S/-Allophone bereits im Sonagramm zu verrechnen. Ein entsprechendes Programm hierzu existiert an der Sektion Informationstechnik der TU Dresden. Hierbei werden mittels dynamischer Programmierung zwei Sonagramme zeitäquivalent übereinander geschoben und dadurch vergleichbar. Diese Methode erfordert jedoch einen hohen Rechenaufwand.

Die Spektralanalyse mit Hilfe der Filterbank dient der Grundlagenforschung und ist nicht für den Einsatz in der Praxis gedacht. Auf der Grundlage der gewonnenen Ergebnisse wird ein System mit Fast-Fourier-Transformation entwickelt, das lediglich den Kleincomputer mit dem ADU-Zusatzmodul nutzt. Die Arbeit mit der Filterbank war jedoch unabdingbar, weil sie gegenüber der Fast-Fourier-Transformation zwei Vorteile bietet: (1) Spektralisierung mit geringem Rechenaufwand und kurzer Rechenzeit, (2) speicherplatzsparendes Zerlegen des Signals bis zu einer Frequenz von 18 kHz (bei 1o bit Auflösung). Bei 64 kbyte des Computers hat die Filterbank eine Grenzfrequenz mit einem Analysebereich von 1.2 s, die bei 2o kHz liegt. Das Zeitfenster bietet einen Ausschnitt von 6 ms.

Das Spektralanalysesystem hat folgenden Aufbau:

Mikrofon → Verstärker → Filterbank Kanäle 1-32

→ ADU → Computer KC 85/3 → Drucker / Externspeicher / Farbdisplay

LITERATUR

/1/ Paul, V.: Modelle für die Verarbeitung fließender Sprache. Nachrichten/Elektronik 37 (1987), 22-23

/2/ Berg, H.: Statistische Untersuchung an einem Spracherkennungssystem. In: 2o. Fachkolloquium Informationstechnik. TU Dresden 1987, 94-98

PROSODIES OF INITIALS IN ENGLISH

BRIAN ANNAN

Department of Linguistics
University of Zimbabwe
Box MP 167, Harare, ZIMBABWE

ABSTRACT

An attempt is made here to illustrate a prosodic analysis of English ('allegro' style), showing that the initial consonants play the major role in determining the pronunciation of English. Some comments on the perception of 'allegro' speech are made which are for further investigation. An insight into the status of liquids in English is presented.

Prosodic analysis as a method of phonological analysis has been well exemplified, mainly through Asian and African languages, but little has been written on European languages, see however /1,2/. The method is basically a top-down analysis which may be completely phonological or phonological explanation for grammatical features /2/. In an analysis of English I have attempted to show how prosodic analysis can make useful generalizations about phonology by starting with the Tone Group (Halliday /3/), the Foot (Abercrombie /4/), the Syllable and the Phonematic Unit. At all these levels abstractions are made called Prosodies and syllable and at the classification of liquids in English. The syllable under each foot consisting of a stressed syllable (plus one or more weak syllables). Please note that in this form of analysis there is no need for a level "Word", although alternative analyses may incorporate it - my approach is basically dealing with speech in an "Allegro" style. In allegro speech the vowels tend to be centralized and thereby lose much of their contrastive features and consonant articulations control to a greater extent intelligibility /5/. The inherent prosodies associated with the initial consonants of stressed syllables thus control the pronounciation of the whole foot. Let us now look at the prosodies of English consonants. Initial consonants in English may be divided into two classes based on phonation, namely Voiced [b d ɡ v ð z ʒ dʒ l r m n] and Aspirated [p t k f θ s ʃ tʃ]. . The so-called semi-vowels [j, w,h] are infact the prosodies associated with syllable structure. Voiced clusters are always Plosive + Continuant [dr-, gl, ɡr, bl, br] ; Voiceless clusters involving plosives have a period of voicelessness on release which is realized by lack of voicing in the continuant, being the equivalent of aspiration in initial plosives. A three-place cluster involves pre-frication in the form of [s-] before plosives which causes the release phase of the plosive to become voiced: it is thus possible to call this aspiration so that all voiceless clusters initially have an exponent of aspiration. Please note that all clusters beginning with [s,f,θ,ʃ] + continuant, the continuant must be voiced, . therefore this justifies my labelling these as aspirates.

It is also possible to divide the initial consonants as to labialization (w-prosody) [θ;ʒ,r,ʃ,ʒ,tʃ,dʒ], , the others having y-prosody [f, v,l,s,z]. If we turn to the problem of the LIQUID in English we see that, in initial position, there is a contrast between clear [l] and [r] labialized and that in clusters they only combine with their appropriate preceding C on the grounds of palatalization, with the exception of [f] which is open for clustering. Note the restriction on [sr-] which must be [ʃr-] and that [l-] signals a foreign feature, e.g. 'Schloer'. There is thus only one phonematic unit [L] in English, especially as, in final position the only possibility is w-prosody.

This Liquid in English we will call /L/ but it is governed by three prosodies, namely [$^{y\ w\ h}$] and we can observe it across different accents of English. So far, we have only dealt with R.P where [L^{y}-] = 'clear' 'l' and [L^{w}-] = [ɹ] and in final position only [-L^{w}]= 'dark' 'l' occur. In intervocalic position [[l] and [ɹ] can occur contrastively and it is here that we have the possibility of [L^{h}] = [ɹ], [ɾ] and [L^{y}] or [L^{w}] dependent on morphological boundaries, although again these differences are reduced in normal speech. /6/. Common speech defects or malarticulations reflect this stance as initial [ɹ] is often replaced by [w] or [ʋ] , namely the w-prosody, and final [l] in monosyllables is replaced by [w] e.g. 'full' [fʊ^{w}] and 'milk' [mɪɔk] where a rounded back vowel shows the prosody. If we now compare other accents of English where e.g. 'post-vocalic' 'r' is pronounced, then in Scots the [l] is always 'dark' with the possibility of non-articulation in syllable-final cf. R.P 'war' [wɔ:] with

Scots 'wall' [wɔ:] : therefore the contrast in Scots in initial position is between $[L^w]$ = [l] and $[L^h]$ = [r].

It has been posited above that there are good reasons for investigating a top-down analysis of speech especially as it provides a different perspective on the relative values of Consonants and Vowels, in that in an analysis of a 'most careful pronounciation' /7/ the characteristics of vowels in 'words' play the major role. We are not dispirting the fact that words have an important role in the learning and production of language, but we are suggesting that, in normal conversation or communication between two speakers, there is a greater dependence on the prosodic features - prosodies - than on words per se. It is my personal experience that perception on a socio-linguistic acceptability level is a question of one listener matching at different levels the performance of a speaker in terms of congruence/lack of congruence: as follows,

a) if intonation is congruent, acceptance: N.B. even if lower levels are observed and,
b) if intonation is non-congruent - check lower levels for possible congruence. If no congruence at any level, then acceptance of non-native speaker of language. This is basically an area which I should like to pursue, especially within the context of event perception /8/.

Brief Outline of a Prosodic Statement of English

TONE GROUP: (Pre-tonic) Tonic$^{*(1-5)}$: Prosodies of pitch contrasts and voice quality.

FOOR: Salient syllables plus weak syllables = $S(W^{1-4})$: prosodies of isochronicity: W syllables are Syllabic C and central unstressed vocoids [ɪ,ʊ,ə] represented prosodically by $[^y, {}^w, {}^h]$

SYLLABLE: Salient syllable = (C)V(C) $^{o/y/w/h}$

PHONEMATIC UNIT: V = phonemic short vowels
C m as described in article above plus some finals not discussed but which mainly are included in the syllable nucleus, namely the vowel.

References

/1/ K.H. Albrow, "Mutation in 'Spoken North Welsh' in In Memory of J.R. Firth" ed. C.E. Bazell et al. (1966)

/2/ K.H. Albrow, "The Phonology of the Personal Forms of the Verb in Russian Archivum Linguisticum, Vol. XIV, Fase.2: 146-156 (1963).

/3/ M.A.K. Halliday, "The Tones of English" Archivum Linguisticum, Vol. XV, Fase.1: 1-28 (1963)

/4/ D. Abercrombie,"Syllable quantity and enclitics in English, in In Honour of Daniel Jones, ed. Abercrombie, D. et al (1964)

/5/ C.A. Fowler, "Coarticulation and theories of extrinsic timing; J. Phon. Vol. 8, 113-133 (1980)

/6/ R.A.W. Bladon & A. Al-Bannerni, Coarticulation Resistance in English /l/, J. Phon. Vol. 4, 137-150 (1976)

/7/ Per Linell, "The Concept of Phonological Form and the Activities of Speech Production and Speech Perception, J. Phon. Vol. 10, 37-72 (1982)

/8/ B. Annan, "Articulatory Base, Coarticulation and Assimilation: some theoretical proposals, XIV Congress of Linguists (1987) forthcoming.

THE CHANGE OF THE PHONOLOGICAL TYPE OF A LANGUAGE

Y. Kuzmenko

Dept. of Indoeuropean Studies
Institute of Linguistics
Leningrad, USSR 199053

ABSTRACT

A pattern of the change from a phonemic language to a syllabic one is established (mora-counting → isochrony → contact correlation → morphosyllabism) on the basis of the evolution of Germanic languages and dialects.

INTRODUCTION

The phonological type of a language, as it is understood here, depends upon the smallest unit of phonological segmentation, which in its turn, is determined by the relationship of the syllable and morpheme boundaries. In the languages where these boundaries do not coincide and the syllable-final consonants may become syllable-initial (cf. Russ. pol - pola with the morpheme boundary after l and the syllable boundary after o: pol-a vs. po-la), units of segmentation less than a syllable can be distinguished (i.e. phonemes). In the languages where syllable and morpheme boundaries always coincide and syllable(=morpheme)-final consonants do not become syllable-initial the smallest unit of phonological segmentation is a syllable[1]. Two types of languages immediatly follow from this segmentation procedure, viz. phonemic and syllabic. One should remember that the notion of syllable is different in each type: while in the former it is purely phonetic in the latter the syllable is a unit of both morphological and phonological level and should be termed therefor morpho-syllable.
There is a third group: languages where these two types of morpheme - syllable boundary relationships coexist and are opposed, namely the languages with the correlation of contact (or syllable cut). The loose contact words are characterized here by the relations typical of phonemic languages (cf. Engl. reader with morpheme boundary after d and syllable boundary after the root vowel), while in close contact words the syllable boundary does not separate the consonant from the preceding vowel (cf. Engl. putting) inherent to syllable languages.
The three types of languages however can represent three stages of the development from a phonemic language to a syllabic one. The stages of this evolution can be found in Germanic languages with their well documented history and deeply rooted tradition of dialectology.

1. PHONEMIC STAGE: FROM MORA-COUNTING TO ISOCHRONY

All Old Germanic languages had free quantitative oppositions of vowels and consonants (CV̆C̆, CV̄C̆, CV̆C̄, CV̄C̄) which did not depend upon the differences in segmentation into syllables, i.e. the syllable boundary could be after a vowel irrespective of its quantity as in Modern Estonian or Lithuanian. Prosodic equivalence of one long syllable to two short ones suggests that in Old Germanic languages the quantity was based on mora-counting, both CV̄C̆, CV̆C̄ and CV̆CV̆ being bi-moric. The situation is similar in some of the modern Swedish and Norwegian dialects where "disyllabic words play the same role in the sentence rhythm as in Early Latin or Old Icelandic: two short syllables being equal to one long"[2].
Most of the root morphemes in Old Germanic languages were bi-moric. According to E.Haugen 75 per cent of Old Icelandic stressed syllables (i.e. root morphemes) were long (i.e. bi-moric) [3]. My own data show that the frequency of mono-moric roots in Old Icelandic did not exceed 12 per cent. Toward the "middle" period monomoric roots in all Germanic languages were ousted by bi-moric ones which resulted in the lengthening of either vowels or consonants in the original CV̆C̆ roots and hence in the equal quantity of all root morphemes (CV:C̆ ∾ CVC:). The ousting of the remaining last mono-moric root signified the end of the mora-counting. The mora-counting correlation transformed into that of syllable length: the law of syllable leveling by which a stressed syllable is always long and consists either of a long vowel (plus a short consonant)

or a short vowel plus a long consonant (or two consonants). Since long and short vowels are possible before short and long consonants respectively the length of each particular phoneme becomes redundant. The situation which may be for convenience's sake termed isochrony was characteristic of all West Germanic languages and of Danish in the "middle" period of their history and is still characteristic of all modern Scandinavian languages except Danish [4,5].
The transition from mora-counting to isochrony has led in all Germanic languages to numerous phonemic changes caused by the elimination of quantitative oppositions (quantity being replaced by quality: V̌ - V̄ > V́ - V̀). Moreover in Swedish and Norwegian the establishment of isochrony has resulted in the transformation of moric peak accents into syllable accents. The change of mora-counting to isochrony does not mean a change in phonological typology for here too the syllable division remains as it is in phonemic languages, i.e. with syllable boundary after vowel (cf. Swedish kärra [čä- r:a] [6]).

2. TRANSITION: FROM ISOCHRONY TO THE CORRELATION OF CONTACT

The next stage in the evolution of the typology of Germanic languages comes when the relevance of the syllable boundary becomes established. This stage which replaced isochrony in all West Germanic languages and in Danish is characterised by the opposition of two different syllable divisions (cf. Engl. pulling - pooling, Germ. Ratte - rate). In close contact words the syllable boundary does not separate the postvocalic (=morpheme-final) consonant from the preceding vowel. In loose contact words the syllable boundary separates the consonant from the preceding vowel in disyllabic words. Two types of contacts and two types of syllable morpheme boundary relationship are opposed here. In Danish the establishment of relevant syllable division was followed by the transformation of syllable accents of Swedish type into the markers of contact. Thus in addition to two types of contact of West Germanic languages we have in Danish two more types of contact (superclose and superloose) effected through the Danish stød. The superclose contact is peculiar to the words with the stød on a consonant (cf. Dan. falder [fal'-ər])while the superloose one to the words with the stød on a vowel (cf. huset [hu'-səđ]) [7]. The superclose contact in CVCV words provides for a complete coincidence of syllable and morpheme boundaries.

Whithin the framework of the correlation of contact there are syntagmatic changes that result in the increased number of words with coinciding syllable and morpheme boundaries which is especially noticeble in English, Low German and Dutch and in still greater degree in Danish with its stød on a consonant (superclose contact). According to L.Brink and J.Lund[8] most frequent changes that occurred in Standard Danish over the past hundred years were the vowel shortening and the shift of the stød, the latter being the change from the superloose contact to the superclose one. The alternations like Danish brev - brevet [bre'v] ~ [breu'] vs. [bre'vəđ] ~ [breu'əđ] reflect various stages of the process leading to the coincidence of the morpheme and syllable boundaries, the forms [breu'] - [breu'əđ] dominating in the contemporary usage. In the languages with the correlation of contact the coincidence of the boundaries and close contact have resulted in the monophonemization of all combinations of vowel plus consonant type which can not be separated by either syllable nor morpheme division, including the combination of vowels and velar nasal. The words like English hang and Danish hæng should be regarded therefore as indivisible morpho-syllables for there is no linguistic procedure of separating the vowels in these words both from the preceding and the following consonant.

3. Syllabic stage: FROM THE CORRELATION OF CONTACT TO MORPHO-SYLLABLE

This change is exemplified by the Danish dialects of Jutland (especially West and South Jutlandic). The number of monosyllables has increased here due to the apocope and the shifting of the postpositional definite article to the preposition (cf. West Jutlandic [ä hu's], [ä kuon] - Standard Danish huset, konen). As far back to the XVIII century the first students of the South Jutlandic observed that "words here are so shortened to look like pure roots" [9]. However the apocope and the shifting of the definite article did not lead to the complete elimination of unstressed suffixal morphemes. The Jutlandic have retained [ə] < [ər] which is the formant of verbs and substantives and the formant of past participle [ən]. However, in most of the forms mentioned root vowel shortening and shifting of the stød have taken place which resulted in the coincidence of the syllable and morpheme boundaries (cf. Jutlandic [diel]-[djäl'ə], [gri'v]-[griu'ə], [kuon]-[kwonə], [brui'-ən], [röi'ən] Standard Danish dele-deler, gribe-griber, kone-koner, past participles of the verbs bryde and ryge). In some Jutlandic dialects even [ə] < [ər] can be apocopated (cf. [fry's], [sgen'], [kwon] - fryser, skinner, koner[10]). The predominance of words in which the syllable and morpheme boundaries coincide has caused the shifting of the syllable boundary in the remaining few disyllabic words with long root vowels (the frequency of such words in dialectal texts does not exceed 2 per cent). In his description of the dialect of Bjerre B.Nielsen points out that here no syllable-initial consonants are possible in unstressed syllables and a long root vowel does not prevent from the coincidence of syllable and morpheme boundaries (i.e. the words of CV'CV or even CV'CCV types have both boundaries after the last consonant of the root morpheme: CV'C-V, CV'CC-V) [11]. It should be noted here that consonants after long vowels are not syllable-initial in the syllabic languages of South East Asia (e.g. in Viet-Namese).
The coincidence of syllable and morpheme boundaries in the Jutlandic dialects has led to the elimination of the correlation of contact and to the change of the phonological type of the dialects.
This change has in its turn caused the change of the function of the stød and the length which are no longer syllable division (and contact) markers but supersegmental features corresponding to the tones of syllabic languages of the Chinese type. Accordingly the dynamic quality of the stød so important for implementing the type of contact and syllable division are less prominent in the syllabic Jutlandic dialects. In the apocope area either a so called weak stød (with dominating tonal component) or solely tonal movement (as in South Jutland) exist. The ousting of the stød by the tone is especially conspicuous in those apocopating dialects that have retained the postpositional article (e.g. in Himmerland). The monosyllables here are characterised by predominently tonal distinctions (cf. [bi'l] with rising-falling tonal movement, Standard Danish bil vs. [bi'l] bile) while the dissyllables have retained the strong stød (i.e. glottal stop) - (cf. [bi'lən] bilen) which is a marker of syllable division [12]. In some of the South Jutlandic dialects the stød and its absence are already ousted by tonal distinctions.
Tones in Danish dialects are believed to reflect an archaic stage, due to their similarity with the Swedish and Norwegian accents [13]. However the data from the dialects of Himmerland and Fyn where the process of ousting is a living one show that tonal distinctions here are secondary as compared to the markers of contact (the stød and its absence). Even in those dialects where the stød retains some dynamic features the coincidence of syllable and morpheme boundaries testifies that it should be regarded as a tone functionally similar to that of syllabic languages of Asia.
The different value of the stød, the length, and the tone in apocopating and non-apocopating Danish dialects was quite clear to P.Andersen who noted that "the phonological function of the length, the stød, and the tone in Jutlandic differs from that of island dialects. The stød in island dialects is a marker of a particular syllable structure, while in Jutlandic it manifests relevant tonemes or accents (or probably even phonemes!)"[14]. Different combinations of the length and the stød (or tonal movements) in syllabic Danish dialects are responsible for five types of syllables (=root morphemes) traditionally termed as tones,Dan. tonehold: CVC, CV'C, CV'C, CVC', CVC'. It is by no means a coincidence that the term tone (tonehold) was coined by the prominent Danish scholar of the XVIII. century J.P.Høysgaard who was of Jutlandic origin. In the syllabic languages of the South East Asia the number of tonal oppositions depends on the quality of vocalic and consonantal elements of the syllable. This is just the case with the Jutlandic dialects where all five tones can occur only in the inviroment of vowel+sonorant, while in the vowel+obstruent position only two types of tones (West Jutlandic stød and its absence) are available. These two tones correspond to so called entering tones of the syllabic languages of Asia.

The similarity between the syllabic languages of Asia and the syllabic Jutlandic dialects is not solely confined to the coincidence of the syllable and morpheme boundaries and to the similar function of prosodic features. The qualitative structure of a syllable in Jutlandic tends to be the same as in the syllabic languages of Asia which is evident from the more intimate juncture of vocalic and consonantal components of a syllable as well as the qualitative and quantitative differentiation of its initial and final components. At the same time Jutlandic differs considerably from modern syllabic Asiatic languages. The syllabic dialects of Jutland are characterised by prosodic and/or qualitative morphonological alternations ([hu's] - [hu's] or [hu's] - [hu's] Standard Danish hus - huse, [söŋ'] - [saŋ'] or [söŋ'] - [saŋ'] synge - sang).
Though in many originally weak verbs the dental suffix is lost (cf. [vän'] - [vän'] vaenne, vaennede - vaennet, [sgei] [sgei'] skedte - skedt) it remains in some verbal forms (cf. West Jutlandic [lok] - [lo?k] - [lot] - [lo?t] luk, lukke, lukkede, lukket or [gløm'] - [gløm'] - [gløm?t] - [glømt] gløm, glømme, glømde, glømt). The frequency of such forms does not exceed 2 per cent , but the sheer fact of their existence may be regarded as the evidence of phonemic segmentation. However, the syllable-final [t] in Jutlandic does not convert into syllable-initial one and this fact does not allow us to

regard [t] here as a separate phoneme. In such forms as [lot] and [glømt] the morphological meaning is indicated by the alternations of the indivisible morpho-syllybles rather than by the phoneme [t]. In phonemic languages the morphological meaning can be signified by a distinctive feature (cf. such morphonological alternations as garsun - garsuin in Irish or lup -lupi in Rumanian). In the same way in the syllabic languages a morphological meaning can be indicated by a distinctive feature of the phonologically indivisible morpho-syllable (its vocalic or consonantal component). Though modern syllabic languages of Asia have now only few examples of morphonological alternations a great number of facultative variants of morpho-syllables here may have resulted from the similar morphophonemic alternations at the early stage of their development [15].

CONCLUSION

To sum up, the evolution of Germanic languages and dialects provides a pattern of the change from a phonemic language to a syllabic one. The pattern involves four stages: mora-counting, isochrony, contact correlation and morpho-syllabism that can be exemplified by modern Danish dialects. This pattern may contribute to foreseeing some trends and shifts that can take place in the Germanic languages (it may be suggested that the next stage in the evolution of Swedish and Norwegian isochrony is the correlation of contact, while in West Germanic languages and in Standard Danish the correlation of contact is to be followed by morpho-syllabism). The same pattern can be employed for the reconstruction of changes that have occurred in the syllabic languages of various families. While mora-counting and isochrony may hardly be consudered as obligatory stages in the languages other than Germanic, the correlation of contact seems to be indispensable as the predecessor of morpho-syllabism.

REFERENCES

I. Гордина М.В. Оразличных функциональных единицах языка. - Исследования по фонологии. Москва, 1966, 177-183.
2. Hesselman B. Huvudlinjer i nordisk språkhistoria. Uppsala, 1952, 247.
3. Haugen E. On the stressed vowel system of Norwegian. - Scandinavian studies presented to G.Flom. Illinois studies in language and literature. 1942, v.29, N 1, 68.
4. Hofmann D. Die "spätgermanische" Silbenquantitätsverschiebung und die Doppelschreibung alter kurzer Konsonanten in den altfriesischen Quellen. - Studia frisica in memoriam Pr. Dr. K.Fokkema. Walters Nordhoff n.v. Grins, 1969, 67-68.
5. Мячинская Э.И. Количественные отношения в фонологической системе среднеанглийского языка. - Вестник ЛГУ, 1985, № 2, 99.
6. Malmberg B. Die Quantität als phonetisch-phonologischer Begriff. - Lundsuniversitetets årsskrift . Lund, 1945, avd. 1, bd 41, N 2, 50.
7. Kuz'menko Yu. Three types of prosodemes in Scandinavian languages. - X. International congress of phonetic sciences. Utrecht, 1983. Dordrecht, 1983, 100.
8. Brink L., Lund J. Dansk rigsmål. Lydudviklingen siden 1840 med saerlig henblik pa sociolekten i København. København, 1975, 221-223.
9. Varming L. De jyske folkesprog grammatisk fremstillet. Kjøbenhavn, 1862, 2.
10. Nielsen B. Et Bjerreherredsmål. - Udvalg for folkemålspublikationer, ser. A, N 23, København, 1968, 23, 44, 52.
11. Nielsen B. Op. cit., 18.
12. Molbaek-Hansen P. Stød and syllabicity in a Jutlandic dialect. - ARIPUC, 1978, v. 12, 16-17.
13. Ringgaard K. Distribution af stød og tonal accent i danske dialektområder. - Nordic prosody, Lund, 1978, 150.
14. Andersen P. Orientering i dansk dialektologi. København, 1954, 85.
15. Fang Kuei Li. Some tonal irregularities in the Thai languages. - Studies presented to Chirô Hattori. Tokyo, 1970, 420- 421.

PHONOLOGICAL TYPE IN MOVEMENT

V.A.VINOGRADOV

Institute of Linguistics
Academy of Sciences
of the USSR

ABSTRACT

Dynamic approach in phonological typology makes it possible to show that the movement of a type in time involves two mutually bound processes: cyclic changes and an outstripping development of the secondary typological features before the primary ones have fully developed.

Many different ways exist for arriving at phonological types, but it is quite impossible to outline them here even briefly. Along with the inventory-configuration approach set up by Trubetzkoy (cf. his typology of vocalic systems), only some of the most typical can be mentioned.

Particular attention paid by Trubetzkoy to configurative traits of sound systems, incidentally, gave rise to purely surface representations of them. Thus, in the not too remote past there was a vogue for "geometric" typology that gave much room for drawing impressive pictures but too little - for enrichment of our still poor knowledge of the deeper structures and languages changes. Quite another, more fruitful trend of the configurative approach concerns the ascertainment of universal implicative relations between the feature categories that set the sound space of a language, cf. /1/.

There was also (and still is) an "arithmetic" typology based on the consonant-to-vowel ratio in sound systems and/or in sound sequences which represent some (no matter how sampled) texts in some (ideally in all) languages. This approach makes it possible to show the rough structure of the phonic substance of expression and to get a phonological typology of rather modest informativeness and of even more modest historical significance. Such a typology is still less promising when it is letters, not sounds that are counted.

Yet another line of typological reasoning deserves closer attention. The phoneme can be considered a mode of storage of the information about the most constant and "pressed" blocks of distinctive features. The other mode can be a syllable, in which case it acquires the status of an emic unit - syllabeme. Thereby a simple but universal typology can be (and actually has been) obtained that distinguishes phonemic vs. syllabemic languages, cf. /2/, especially /3/. This approach has the merit of being not eurocentric since it is applicable to practically all languages.

All the typologies mentioned above, however, are obviously static by their nature, whereas more effective is a dynamic approach, most soundly vindicated by J.H.Greenberg, aimed not merely at the arrangement, but at the explanation of the attested data by means of a complex procedure including both synchronic and diachronic argumentation (cf., for instance, /4/).

Such an approach implies the focusing of attention on the evolution of phonological types. There are two, to a certain extent mutually exclusive, properties of any language (permanent changeability and restrictedness of language technique) which induce one to suggest that the main principle of the language-type movement in time should be a cyclicity, despite the fact that some particular changes in the system can be irreversible. The cyclicity is apparent both in separate cells of the system and in the whole inventory of its items. It must be noted, though, that the cyclic movement by no means implies a precise repetition of the previous state of the system (type). In other words, a chain of evolution represented as $S_o \rightarrow S_i \rightarrow S_o$ (where S_o is some initial state) is but a particular case along with a more common case: $S_o \rightarrow S_i \rightarrow S'_o$, i.e. the question is not of a return to the initial state but of the transition into a state partly mapping the previous one.

Thus, in Polish a current process of denasalization of the nasal vowels is observed, but it leads to the restitution of the "initial" state only in some environments. There are yet other positions where the nasality either disappears ($\tilde{V} \to V$) or is substituted by the labiality ($\tilde{V}$ -- Vų), e.g.zęby -- /zemby/ 'teeth', but idę -- /ide/ 'I'm going', z tobą -- /stoboų/ 'with you'.

An example of the cyclicity at the level of a whole subsystem can be found in Bantu languages where some of them have seven-phoneme vocalism while the others have five-phoneme vocalism. As soon as for the Proto-Bantu a system of seven vowels has been reconstructed while for the Proto-Benue-Congo (of which the Bantu is an offspring) - a five-vowel system, it means that we are faced with a cyclic movement: *5 → *7 → 5.

The above examples present simple and evident cases of cyclicity. One can find more complicated instances as well where an analysis of the phonological movement implies a phonological type to be correlated with more general types (in the first turn, with grammatical types).

In these latter cases another peculiarity of the type evolution can come to be better noticeable - an outstripping development of the secondary (implied) traits while the primary (type-forming) features have not yet evolved as a complete system. Such an evolutionary situation could be called the type anticipation, but in reality not every fact of the outstripping development means an obligatory transition to a new type: the movement can slow down or completely cease, and the new secondary features already in view either vanish or undergo restructuring. Below, a probable example of a "deep" phonological cyclicity will be regarded in some details, which is bound up with the evolution of the hierarchically higher - morphological - type.

Generally speaking, in every language the phonetic system may contain some more or less evident discrepancies in the sound production (or variation), due to the articulatory nature of different sounds or to a change (drift) affecting the system. In the last case, those discrepancies could be of a high typological value as indicative of some profound changes that affect the phonological system as a representative of a certain phonological type. However, within the Procrustean bed of rigorous phonemic interpretations, most of the allegedly marginal phonic anomalies inevitably come to be smoothed over. Undoubtedly, a functional view of speech data is a virtue and the very essence of linguistic analysis, but this view must not lose sight of the incessant movement of "language matter".

In the Russian phonetic system, the following discrepancies and anomalies can be observed that are of interest in view of the topic in question:

(1) acoustically poor and articulatorily non-homogeneous expression of the non-privative opposition of the "hard/soft" consonants mostly identified according to the properties of subsequent vowels /5;6/;
(2) weakening of the rule of assimilative "softing" of consonants before a soft consonant, the process spreading from morpheme-juncture positions to the intramorphemic positions /7/;
(3) a much higher degree of the articulatory cohesion in the unstressed syllables as compared to that of the stressed ones /8/;
(4) a sort of harmonization of the unstressed syllables /9/;
(5) phonetic asymmetry of the syllable patterns CV/VC, the latter showing the extraphonemic elements of Knacklaut type before V- and a voiceless vowel after C- /10/.

These apparently strange facts can be clear enough if put into relation with the agglutinative tendency observed in the Russian morphemics /11; 12/, i.e. with a change of the morphological type, and then (2) and (4) could be treated as facts of the outstripping development of secondary syntagmatic (antifusional) traits of a new paradigmatic type that is but taking shape. As regards the (1), (3) and (5), they reflect perhaps the competition of the cyclically bound phonological types - phonemic and syllabemic, the (3) probably meaning that in the Russian phonetic word two alternative patterns coexist - the phonemic one in stressed syllables and the syllabemic one in the stressed ones. The syllabemic type is supposed (R.Jakobson, R.I.Avanesov, V.K.Žuravlev) to be older, therefore we must be faced with the cyclic movement "syllabemic type -- phonemic type -- (quasi)syllabemic type". At the present stage, the elements of the syllabemic type turn out to be consonant with the new tendency to the agglutination proper to the Russian morphemics.

A phonetic background of the cyclicity outlined here could be seen in the far gone qualitative variation of vowels due to the consonantal domination in the Russian syllable, which causes the phonetic unstability of morphemes. And what is more, the usurpation of the timbre category ("fronted/backed") by consonants threatens the very existence of vowels as an autonomous phonemic class, that is why in the discrepancies mentioned above one can see a "phonetic riot" against phonemic regime, a riot directed towards the morphemic process of restructuring that goes from above.

REFERENCES

1. J.Crothers. Typology and Universals of Vowel Systems. In: J.H.Greenberg (Ed.) Universals of Human Languages. Vol.2: Phonology. Stanford UP, 1978, p. 93-152.
2. V.B.Kasevič. Fonologičeskije problemy obščego i vostočnogo jazykoznanija. Nauka, Moskva, 1983.
3. Kao Xuan Hao. Phonologie et linéarité. Reflexions critiqes sur les postulats de la phonologie contemporaine. SELAF, Paris, 1985.
4. J.H.Greenberg. Rethinking linguistics diachronically. "Language", 1979, vol. 55, N. 2, p. 275-290.
5. L.R.Zinder, L.V.Bondarko, L.A.Verbickaja. Akustičeskaja xarakteristika različija tverdyx i m'agkix soglasnyx v russkom jazyke. - In: Voprosy fonetiki. Učenyje zapiski Leningradskogo universiteta, N. 325, serija filologičeskix nauk, vypusk 69, 1964, s. 28-36.
6. A.A.Reformatskij. O korrelacii "tverdych" i "m'agkix" soglasnyx (v sovremennom russkom jazyke). "Cercetări de lingvistică", anul III, 1958, Supliment, p. 403-407.
7. M.L.Kalenčuk. Osobennosti realizacii soglasnyx fonem na styke morfem v sovremennom russkom literaturnom jazyke. Avtoreferat kandidatskoj dissertacii. Moskva, 1986.
8. L.V.Bondarko, L.A.Verbickaja, L.R. Zinder. Akustičeskije xarakteristiki bezudarnosti (na materiale russkogo jazyka). - In: Strukturnaja tipologija jazykov. Nauka, Moskva, 1966, s. 56-64.
9. R.F.Paufošima. Fonetika slova i frazy v severnorusskix govorax. Nauka, Moskva, 1983.
10. L.V.Bondarko. Struktura sloga i xarakteristika fonem. "Voprosy jazykoznanija", 1967, N.1, s. 34-46.
11. Slovoobrazovanije sovremennogo russkogo literaturnogo jazyka. Nauka, Moskva, 1968.
12. Fonetika sovremennogo russkogo literaturnogo jazyka. Narodnyje govory. Nauka, Moskva, 1968.

SPATIAL CONFIGURATION OF TYPES OF PHONOLOGICAL SYSTEMS OF CENTRAL AND SOUTH-EUROPEAN LANGUAGES

M.I.Lekomtseva

Institute of Slavic and Balkan Studies of Academy of Sciences of USSR. Moscow

ABSTRACT

The paper presents an areal classification of the phonological systems of 18 languages of Central and South-East Europe. By means of cluster analysis of phonological specifications of these languages two areal types are obtained: Balto-Balkanic type, presented by two areas, and the Central type which is situated between these areas.

After the discovery of phonological features N.Trubetzkoy and R.Jakobson were led to conclude that the distribution of phonological features in different languages was not random. The areal distribution of the phonological features resembled the isoglosses of lexemes studied in linguistic geography. At the 1st International Congress of Linguists N.Trubetzkoy outlined the task of studying the areal configurations of linguistic features /9/. R.Jakobson presented the first example of as language area (the Euro-Asian Sprachbund or linguistic area) modelled on the basis of compact territorial distribution of languages of different degrees of genealogical relation. The phonological systems of the corresponding languages were characterized by the two following dominant features: palatalization of consonants and monotony of vowels /4/. Areal language studies show that the phonological system of a language derives not only from its genealogical relations, but is likewise influenced by the neighbouring languages /2, 3, 5, 6, 7/. The development of the phonological system of a language is further conditioned by the language situation in the given region and the spatial structure of the linguistic area.

While the task of finding out the distribution of phonological features is allover clear, the choice of the guidelines for setting up a basis for defining the areal type of phonological systems (i.e. the domain of connectedness of a linguistic area) presents serious difficulties /3, 6/. E.g. though the phonological system of Irish is characterized by the two abovementioned dominant features: the opposition of palatalized/nonpalatalized consonants and monotony of vowels, the Irish language can hardly be recognized as belonging to the Euro-Asian language area. Two coincidences of features are not enough to decide whether the language belongs to a definite type /against: 3/.

The Balkan language area (like other currently defined language areas) does not hold such features that might be regarded as necessary and sufficient in the traditional sense. It is frequently noted in areal typology that every feature defined as specific for a linguistic area can also be found in a language beyond the area /3, 7/. On the other hand, some languages belonging to the area may lack a feature defined as specific. Essentially it is a high degree of similarity characterizes the languages belonging to the same domain of connectedness of a linguistic area /1, 2, 3, 6, 7/. The setting up of abstract ideal types to serve as a basis for a quantitative evaluation of real language systems /3/ seems to be of little value: the problem becomes one of constructing ideal types to be used for a further description of the language systems of a given area. A high degree of similarity among the languages of a given area however is frequently due to non-exclusive intersecting features, that do not fall into clear-cut patterns.

This paper presents an attempt of suggesting areal types of phonological systems on the basis of the languages of the Baltic-Balkan areal.

The phonological systems are viewed with regard to their inherent features, i.e. the features of monotony - polytony go beyond the scope of our study. Phoneme identification was accomplished according to the distinctive features of Chomsky-Halle and their amendment in Halle-Stevens. The syntagmatic aspect of the phonological systems was not taken into account.

Areal studies naturally fall into the domain of dialectology. N.Trubetzkoy regarded them as a continuation of dialectal studies /9/. The language material used and the way we have set our task inclines us, however, in the present preliminary stage of analysis to consider the area as represented by languages in standard form. Further tasks include a description of dialects as global phonological systems, each with a given set of features and setting up areal types of dialectal phonological systems of genealogically related and unrelated languages.

The areal types are set up on the basis of the phonological systems of the following languages belonging to the Balto-Balkan linguistic area: Latvian, Lithuanian, Upper-Lusatian, Polish, Belorussian, Russian, Ukranian, Slovak, Check, Hungarian, Romanian, Bulgarian, Turkish, Greek, Albanian, Macedonian, Serbo-Croatian and Slovenian languages.

A transparent procedure of construction, controlled objectivity on all stages and unambiguity of results are essential for setting up areal types of phonological features.

Due to the abovementioned peculiarities of areal language relations the traditional Aristotelean classification is believed to be of little value. In the present case a quantitative approach seems preferable. Of all techniques of cluster analysis, employed for the purpose of obtaining groups of objects characterized by a maximum degree of similarity, we have considered the Linker algorithm as best suited for our needs /8/.

The Linker algorithm gives hierarchically arranged object clusters and determines the relative degree of similarity by which the objects are clustered. Generally the algorithm guarantees a local rather than a global optimum. In our case if the objects can be enumerated, the algorithm guarantees the global optimum as well /8/. Consequently, if the data matrix has inherent structure cluster analysis will succeed in identifying it. The result of the Linker algorithm is unambiguous only if the degrees of similarity (or distance) among the objects are different from each other. If, however, there are identical degrees of similarity between two or more pairs of objects (as may frequently be the case in areal studies) the result of clustering becomes ambiguous. To eliminate this disadvantage of the Linker algorithm we have introduced a subalgorithm to be applied in the case of identical degrees of similarity (or distance). The subalgorithm assigns a clustering value to each claimant depending on the next step of the main algorithm, i.e. the joining of a given claimant with all the other clusters is preferred if this results in the maximum sum of the mean degrees of similarity (or correspondingly the least sum of the mean degrees of distance).

The Linker algorithm can be applied to the matrix of distance (or similarity) between given languages if metric space has been determined. The condition of metricity is met by applying the formula defining distance /8/ between the phonological systems of the given languages. I.e.

$$d = 1 - \frac{\alpha + \delta}{l}$$

where

$$\alpha = \sum_{k=1}^{l} \min(x_k, y_k)$$

$$\beta = \sum_{k=1}^{l} x_k - \alpha$$

$$\gamma = \sum_{k=1}^{l} y_k - \alpha$$

$$\delta = l - (\alpha + \beta + \gamma)$$

l - the number of positions (features) chosen to represent the given language. Else: α - the number of positions (features) where both languages have a positive values (1); δ - the number of positions (features) where both languages have a negative value (0).

The data matrix of the distance between the given languages is filled according to the abovementioned formula.

According to the algorithm the least distance between the languages is selected and the corresponding languages clustered. Next the mean distance from the obtained cluster to the rest of the languages is calculated. The languages showing minimum distance once more undergo clustering. In the case of several identical distances we introduce the subalgorithm. The routine is run until all the languages have been clustered.

The algorithm can be presented in the form of a dendrogram mapping the sequence of element and group clustering and showing the minimal distances at which the clusterings take place.

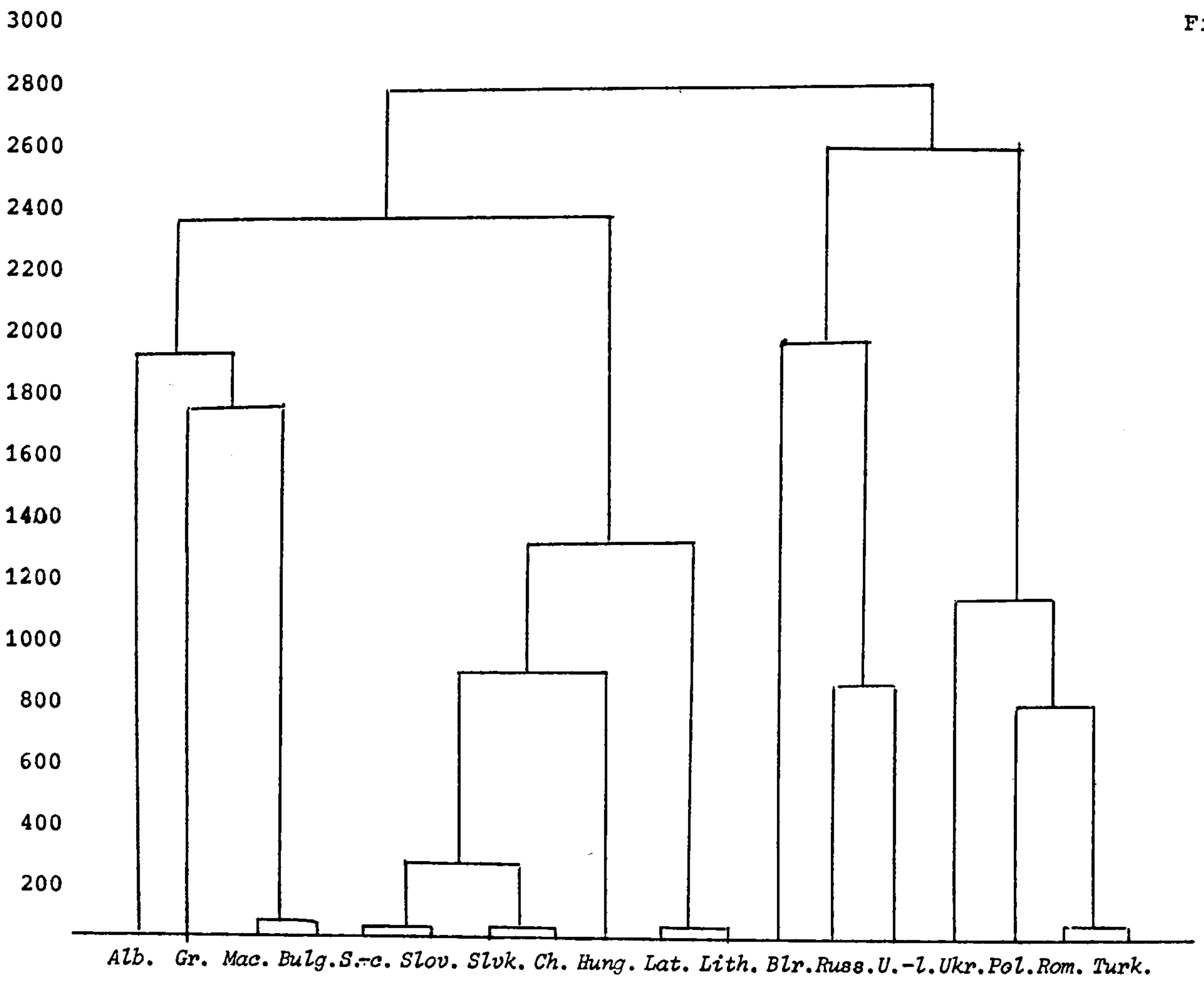
The languages considered present the following picture of language groups marked by an increasing degree of similarity in paradigmatic phonology (see Fig. 1).

Fig. 1 shows the maximum similarity (minimum distance) between the phonological systems of Romanian and Turkish (similarity, here, is viewed paradigmatically, - f.e. the introduction of syntagmatic features would naturally change the position of both Romanian and Turkish). Likewise the maximum degree of similarity characterizes Latvian and Lithuanian, Check and Slovak, Serbo-Croatian and Slovenian, and Bulgarian and Macedonian.

The next cluster is formed by Check, Slovak, Serbo-Croatian and Slovenian, while the following step adds Hungarian to the cluster. Last the group including Latvian and Lithuanan is added.

The Bulgarian-Macedonian cluster is joined by Greek, and next - by Albanian. It should be noted, however, that the distance between the initial language cluster (Bulgarian-Macedonian) and the Greek is much greater than the distance between the languages of the formerly mentioned language group. Both language clusters are united into one which is globally opposed to another cluster formed by consequent joining of Uk-

Fig. 1



ranian (Polish (Romanian Turkish))) and (Russian-Upper Lusatian) Belorussian.

It should be noted that while the clustered languages demonstrate a high degree of similarity, none of them is characterized by one or several of specific features. One and the same feature could appear as almost specific or almost non-specific depending on the concrete constellation of features in the given language and its neighbours. Thus the opposition of palatalized vs. non-palatalized consonants was noted to be specific not only to the languages of the first cluster, but also to mark the phonological systems of the languages belonging to the 2nd group (Lithuanian and Bulgarian). This, however, did not interfere with their clustering correspondingly with the Latvian language and the Macedonian language on the basis of whole sets of features.

The dendrogram showing the clustering of the phonological systems depending on the degree of their similarity can be represented on the map as a structure of the corresponding linguistic territory. Thus the relative similarity of languages can be shown as an altitude of the corresponding territories and marked by different hatching. In the present case we have established two areas of similar phonological system or two areal types: Balto - Carpathian - Balkanic area represented by two territories that are separated by a compact Central area of phonological systems including languages from Upper Lusatian to Russian.

The features of syntagmatic phonology as well as features of other language levels should be likewise included in the data matrix. The interpretation of areal types thus obtained will yield the best results when an increased degree of approximation of the global characteristics of the language does not change the output of the algorithm.

CONCLUSIONS

1. From the point of view of paradigmatic phonology the Balto-Balkanic region is a single relatively connected linguistic area with two language areas: a marginal Balto-Balkanic area proper and a central compact area, separating the two territories of the Balto-Balkanic language area.

2. A quantitative analysis of the degree of similarity characterizing the phonological systems of a given area helps to establish the systems of maximum similarity, the hierarchy of their clustering and the optimum of the borderlines between the obtained areal types.

3. The areal types obtained as a result of the cluster analysis described in the present paper show a high degree of overall similarity, while they do not necesserily contain specific features. As specific features characterizing areal types tend to be absent, the procedure of establishing the types is believed to be rather objective.

REFERENCES

/1/ L.Campbell, T.Kaufman and T.Smith-Stark. Meso-America as a linguistic area. - Language, v. 62 N 3 (1986), 530 - 570.

/2/ M.B.Emeneau. Language and linguistic area. Stanford, 1980.

/3/ H.Haarmann. Aspekte der Arealtypologie. Die Problematik der europäischen Sprachbünde. Tübingen, 1976.

/4/ R.Jakobson. К характеристике евразийского языкового союза. Paris, 1931. Reprint: R.Jakobson (1971). Selected Writings. I. Phonological Studies. The Hague-Paris, 144-201.

/5/ R.Jakobson. Über die phonologischen Sprachbünde. - TCLP, IV (1931), 234 - 240. Reprint: R.Jakobson (1971), 137 - 143.

/6/ R.Kattein. Zur Definition des Begriffs Sprachbund. - Sprachwissenschaft, B. 11 H. 3 (1986), 276 - 287.

/7/ C.Masica. Defining a linguistic area: South Asia. Chicago, 1976.

/8/ H.Späth. Cluster-Analysen-Algorithmen zur Objektklassifizierung und Datenreduktion. München Wien 1975.

/9/ N.S.Trubetzkoy. Phonologie und Sprachgeographie. - TCLP, IV (1931), 228 - 234.

ENGLISH IN THE CONTINUUM OF SYLLABIC -- NON-SYLLABIC (PHONEMIC) LANGUAGES (RP AND DIALECT DATA)

OLGA BRODOVICH

Dept. of English Philology
Leningrad University
Leningrad, USSR, 199034

ABSTRACT

Analysing linguistic facts observed in RP and English dialects, the author presents evidence suggesting that, contrary to the established views, English typologically occupies a place intermediate between purely non-syllabic and syllabic languages.

The predominant concept of the role of the syllable in English is that it is a purely articulatory phenomenon, viz., an articulatory unit. But there are facts related to the phonological structure of English words, and especially facts of dialect variation, that are hard to accomodate with this established view.

Let me first point out the fact that /i/ in feat is in most dialects, including RP, much shorter than /i/ (apparently the same phoneme) in lead. That is only one example of the by now well-known feature of the dependence of the English vowels, not on their own place in the short--long dichotomy, but rather on the type of the syllable-final consonant - an example of the re-evaluation of the role performed by the syllable, currently developing within the structure of English.

There is also another angle to the problem. Had the English phoneme been as independent a unit as is suggested by the concept of the minimal unit of surface structure, words like feeling and Ealing would have had one and the same syllabic structure - and that throughout the entire English-speaking world. However, facts of dialect variation show that there is an important difference in syllabic structure between words like feeling, whose morphemic structure is feel+ing, and words like Ealing, with no morphemic boundary. The difference in question consists in that some dialects develop an [ə]-glide before the /l/ in the feeling words, whereas no such thing happens in the Ealing words /1/. It is plainly the consequence of the /l/ in feeling being the dark [ɫ], while in Ealing, the clear [l]. That means that in feeling /l/ belongs to the first syllable, while in Ealing, to the second. This is one of the facts that show that in English, wherever a morpheme boundary occurs after a consonant, it tends to be also the place of a syllable boundary, the final consonant of the morpheme tending to remain syllable-final even when a vowel-initial morpheme is affixed to it. Although this fact has often been mentioned in the literature, it seems that its implications for the structural role of the syllable in English have so far escaped the notice of theorists.

Nowadays our theory is benefited by an important contribution by Prof. Vadim B. Kasevich whose profound re- examination of the syllable, its structure and functional role in various languages enabled him to come up with an entirely new, indeed, a revolutionary system of language typology. According to this theory the world's languages form a continuum with two extreme types - the ideal phonemic type, the ideal syllabic type and a number of intermediate types and sub-types /2/. Should English belong to the first of these extreme, or polar, types, such facts as those described above simply could not happen, for in purely phonemic languages syllabic structure of the words is entirely independent of their morphemic structure, thus freeing the phonemes of any dependence on their place in the syllable.

A careful examination of the facts of English phonology shows that many of these have most striking analogies in syllabic and near-syllabic languages. The results of an attempt at such an examination will be reported in my book due to appear in print in Leningrad University Press in the near future/3/. What follows is a concise version of the analysis.

An important division between purely syllabic and non-syllabic languages is formed by the existence/non-existence of non-syllabic morphemes. In syllabic languages morphemes that do not form a syllable are impossible. In English, it would seem, we do find such non-syllabic morphemes, for such are the morpheme -s (-s^1: 3rd p. singular, present tense of verbs; -s^2: plural of nouns; -s^3: the possessive); the morpheme -t/-d (past tense); the morpheme -th (ordinal numbers) and, finally, the morpheme -th/-t (non-productive suffix of abstract nouns, as in length).

But all of these (except the non-productive -th/-t) have syllabic allomorphs: looks - kisses; books - faces; man's - Jones's; looked - wanted. The importance of the fact of non-syllabic morphemes having syllabic allomorphs in determining the place of the language in the above-mentioned continuum has been amply proved in the literature /4/. It should also be noted that the non-syllabic morphemes are only a very small part of the morphemic repertoire of English. Moreover. Whereas this very modest number of non-syllabic morphemes does actually exist, there are no non-syllabic words in English. Compare this to such a typically phonemic language as Russian, where many prepositions are non-syllabic.

An important feature of syllabic languages is the impossibility of what Prof. V.B.Kasevich terms "re-syllabification", i.e. the process of shifting a syllable boundary from, say, a position after a consonant to a position before the consonant on adding a vowel-initial morpheme to the previous structure. Cf. the Russian examples of dom # -- do # ma. Now, we have already mentioned the fact that Modern English shows a considerable resistance to this process, retaining the syllable boundary wherever possible in the place of the morpheme boundary. This feature of English has also been uncovered by experimental investigations /5,6/.

Another typological characteristic of syllabic languages (one that is closely linked with the above-mentioned) is the predominance of the (C)VC syllable type in such languages, almost to the exclusion of the CV syllable type. Now, English shows a clear trend towards the (C)VC type of syllable forming the most frequently occurring type. Experimental results reported in /6/ are most persuasive. The preference of the English to place the syllable boundary, not before a consonant, but after it has been well proved by B.Malmberg in experimenting with nonsense sequences, such as ipi, opo, apa /7/.

In syllabic languages the vowel is closely linked to the following consonant and is fairly independent of the previous, a fact which is related to the above-mentioned composition of the syllable. In English we observe many instances of the vowel being closely connected with the following - and not the preceding - consonant. If a vowel's variation is subject to any influence of a consonant - it is always the following consonant, and not the preceding, which exercises this influence. Examples are numerous. The realization of historical /ă/ as /a:/, and not as /æ/, occurs in RP and in the South in general before a certain group of consonants, as /s, θ, f/ and sometimes /n + C/. In the case of historical /ŏ/ a long vowel also develops before the same consonants, but more frequently in the North. In many southern dialects, including RP, /æ/ is lengthened before /d/ and some other voiced consonants /8/. Cf. also the well-known development of historical short vowels before /r/ in r-less dialects. The influence of etymological /r/ is now being mirrored in the way vowels are influenced by the following /l/ (cf. [fiəliŋ] mentioned above). Many illuminating examples of the influence of the syllable-final /l/ on the preceding vowel were cited by Dr.P. Trudgill. Thus, in his investigation of the development of East Anglian correspondences to the RP /əʊ/, P.Trudgill found out that the process of approximation of the /u:/ in no and /ʌu/ in know into a form of [ʊʊ] is hindered by the following /l/, where no movement towards a more RP-like sound is noted /9/.

A very important feature of syllabic languages is the difference in ways of variation (diachronic and synchronic) of syllable-final and syllable-initial allophones of consonants. In this, English shows many instances of similarity with syllabic languages. In RP, for example, only syllable-initial /p,t,k/ have aspiration, whereas only syllable-final /t/ is glottalized or (again only in syllable--final position) a glottal stop develops before a group of voiceless stops. It was only syllable-final or syllabic /r/ that was vocalized in the so-called non-rhoting, or r-less dialects. And it is precisely the same position where /l/ is now being vocalized in London speech and in a number of counties nearest to London, especially to the south-east of London. The realization of /ð/ as either [d] or [v] (London) depends on whether it is syllable-initial ([d]) or syllable-final ([v]). Examples where the variation of a consonant is independent of its position in the syllable are very few. Among them are affrication of /t/ or the realization of /θ/ as [f] in London speech.

To sum up. There are many features of the English syllable demonstrating that in Modern English it is something more than merely an articulatory unit. These features are:

- the scarcity of non-syllabic morphemes and the fact that all of them possess syllabic allomorphs;

- The trend for a morpheme boundary to coincide with the syllable boundary;
- the resistance of English to re-syllabification;
- the close contact of vowels with the following consonants;
- the dependence of vowel variation only on the following consonant;
- the difference in variation patterns of syllable-initial and syllable-final allophones of consonants.

None of these facts are entirely new to theorists. But their typological importance, it seems, has been overlooked. These facts have been discussed in the literature in connection with different theoretical problems. Considered together, these facts show that the syllable in English is developing into a peculiar unit of surface structure, somewhere in between the phoneme and the morpheme, and that it is moving in the direction of coalescence with the morpheme. The purely asemantic syllable of the non-syllabic, phonemic languages is being gradually ousted by the syllable which is typically a morpheme.

To be sure, English greatly differs from such purely syllabic languages as Vietnamese or Chinese, but it shows in many ways a trend to develop into a type intermediate between non-syllabic (phonemic) and syllabic languages.

References

/1/ J.C.Wells. English Accents in England. In: P.Trudgill, ed. Language in the British Isles. Cambridge, 1984.
/2/ V.B.Kasevich. Phonological Problems in General and Oriental Linguistics. Moscow, 1983. (In Russian).
/3/ O.Brodovich. Aspects of the Theory of Dialect. Leningrad; in print. (In Russian).
/4/ A.K.Ogloblin. Diachrony and the Morphonology of Malay-Javanese Languages. "Voprosy Jazykoznaniya", 1985, No 3. (In Russian).
/5/ R.K.Potapova, N.G.Kamyshnaya. Syllabification as Viewed from Segmentational Speech Function. "Voprosy Jazykoznaniya", 1973, No 3. (In Russian).
/6/ A.A.Pedersen. Comparative Analysis of English and Danish Syllable Structures. Candidate thesis. Abstract. Moscow, 1980. (In Russian).
/7/ B.Malmberg. The Phonetic Basis for Syllable Division. "Studia Linguistica", vol.9, 1955, No 2.
/8/ J.C.Wells. Accents of English, vol.2. Cambridge, 1982.
/9/ P.Trudgill. On Dialect.Social and Geographical Perspectives. Oxford, 1983.

SPEECH RHYTHM
(main approaches and definitions)

ANTONINA M.ANTIPOVA

Department of English Phonetics
Maurice Thorez Moscow State Institute of Foreign Languages
Moscow, USSR 119034

ABSTRACT

The report sets out a brief review of main trends and concepts on rhythm.

Among the numerous definitions of rhythm two main ones can be singled out: 1. rhythm is an alternation of contrastive speech events (usually stressed and unstressed syllables); 2. rhythm is a periodicity of similar and isochronous (on the perception level) events.

At first glance, these definitions may seem to contradict each other. In reality, they only accentuate different aspects of the same phenomenon.

Those phoneticians who concern themselves with the study of a text usually look upon rhythm as a hierarchical system consisting of units of different size and value. In this connection two main questions arise:

1. What units can form periodicities?
2. What speech segments (syllable, rhythmic group, sense-group, phrase, or supraphrasal unit) can perform the function of a rhythmic unit?

In speech, there can be an alternation of sounds, syllables, sense-groups (tone-groups), phrases (utterances) and supraphrasal units of different types. In verse lines and stanzas can also alternate. Alternating elements are opposed to each other, this opposition being based on different features. On the segmental level, vowels are opposed to consonants. This opposition is based on the presence or absence of noise. This type of alternation is characteristic of languages with syllable structure CV. In English this type of alternation has infrequent occurrence. Only occasionally in poetry does a syllable become a rhythmic unit.

On a higher level, the opposition is based on the degree of prominence, i.e. stressed syllables alternate with unstressed ones. In English this type of alternation is more regular. Stressed and unstressed syllables form a unity which shows periodicity. This unity is usually referred to as a "rhythmic group" (an accentual group). It often coincides with a word. In English a rhythmic group can be considered to be the smallest basic rhythmic unit as it occurs both in prose and in verse.

The next segments larger than the syllable are the sense-group, the phrase and the superphrasal unit. The aforementioned segments can alternate with a pause. Here the opposition is based on the presence or absence of phonation. In this respect the mentioned segments exhibit a different behaviour. Sense-groups are not always separated by pauses, so the "phonation-pause" alternation is not obligatory. A phrase alternates with a pause more frequently than a sense-group. Supraphrasal units, as a rule, is separated by pauses.

A sense-group is formed by lexico-syntactical and prosodic means. In the case of lexical repetitions and parallel constructions, sense-groups are perceived as similar in structure. If lexico-syntactical means vary, their similarity is based on prosody. The beginning of a sense-group is usually marked by the maximum pitch and intensity and a slower tempo compared to the central part of the sense-group. The end is marked by the minimum pitch and intensity, often by the falling tone (in the author's material 84%), and by a slower tempo. The body of a sense-group is characterised by a descending pitch contour (regular or irregular depending on the speaker, style, emotions,etc.) The majority of sense-groups contain 2-4 stressed syllables the total length not exceeding 2-3 seconds, the most typical length being 1-1.5 seconds. Thus all these means make sense-groups similar /8/. At present linguists are researching the alternation of different types of sense-group /18/.

It appears that semantically dominated sense-groups form rather regular periodicities and alternate with semantically neutral sense-groups, just as emotionally coloured sense-groups alternate with emotionally neutral ones. It has also been observed that semantically dominated sense-groups tend to occur in marginal areas of supraphrasal units in prose and of stanzas in verse /12, 17/.

Supraphrasal units are characterized by the same prosodic means as sense-groups but the number of features which characterize the individual supraphrasal units decreases whereas the isochrony increases. Supraphrasal units alternate with pauses and form the "S.Ph.U + pause" complex, which is periodically repeated. A long phonation period is frequently followed by a short pause or vice versa. In other words, a pause here can perform an equalizing function.

Phrases occupy the intermediate position between sense-groups and supraphrasal units. They alternate with a pause more frequently than sense-groups and less frequently than supraphrasal units. Considerable variation in length and rather a vague prosodic similarity prevent phrases from functioning as frequent rhythmic units. They play this role when they coincide either with a sense-group or with a supraphrasal unit.

Thus, practically all speech segments can function as rhythmic units if they become isochronous and similar in character.

In dialogical speech, in addition to the aforementioned types of alternation, the alternation of cues (the speech of each interlocutor) can be added. If a pair of cues (a stimulus and a response) is more or less isochronous they are normally perceived as periodic units. The phonation period (period between two pauses) can form fairly regular periodicities. Periodicities can also be formed by a phonation + pause period, by a series of falling tones, and even by hesitation pauses. Thus, not all the units in spontaneous speech form a hierarchical system. Apparently, the rhythmic system is of a more complicated nature /13/.

Consequently, periodic events can consist of contrasting (in the case of alternation) and non-contrasting elements (in the case of sense-groups). If there is a regular alternation of elements, the leading role in the regulation of rhythm is performed by time, i.e. isochrony. If the alternation is not regular, the leading role in forming periodic units is performed by accentual and tonetic features.

Thus, the perception of speech elements as periodic is determined by the

two factors: isochrony and qualitative similarity. The latter can be expressed by a contrastive complex, (alternation), or by a system of accentual and tonetic characteristics.

Rhythm has been defined as a periodicity of similar and isochronous events, Isochrony can be of two types: physical, which presupposes physical identity of intervals, and perceptual, which presupposes similarity of intervals on the perception level. This last assumption is largely based on works of phsychologists. As it comes from numerous experiments concerned with the perception of intervals of different size (different within certain limits), variations in length can be ignored and physically different intervals can be perceived as similar. There is apparently a process of mental equalization at work /1, 3/. According to some experimental data, non-verbal rhythm (the intervals between recorded clicks) is perceived as stable, with as much as 14.5% displacement of temporal regularity.

Data pertaining to the perception of speech rhythm vary greatly, but there is some evidence to suggest that the size of units (intervals) perceived as regular is relevant: the larger the unit, the greater the difference in length that can be ignored /5, 6/.

A retrospective look at the studies of rhythm gives an idea of how the linguistic approach to this phenomenon developed.

In the 1920s and 1930s Russian linguists conducted extensive textual research. In particular, poetic rhythm was regarded as a hierarchy of rhythms /13,19/.

In the 1950s and 1960s - a period of structural and generative views, when an utterance (not a text) was at the centre of linguists' attention - rhythm was normally understood as an alternation of stressed and unstressed syllables /2, 4/.

In the 1970s and 1980s - a period of close attention to textual problems -, rhythm came to be understood as a system again. By that time, many facts pertaining to rhythm had been accumulated by different sciences (primarily, biology and physics), which gave grounds for considering rhythm to be a fundamental law of the structure and development of the material world.

Achievements in the development of philosophical ideas as to the character and structure of a system largely contributed to the study of speech rhythm. In 1974 the authors of the book "Rhythm, Space and Time in Literature and Art" actually raised the problem of unity of these fundamental forms /14/.

Rhythm, being a periodicity, organizes events. It organizes the space-time continuum and the events themselves. Rhythm can be regarded as a general language system that organizes a language as a whole. A large proportion of rhythm research is concerned with the linguistic form investigated in the context of the meaning it conveys. Views on rhythm as a functional unit are characteristic of Russian linguistics. The works of A.M.Peshkovsky /13/, B.M.Eihenbaum /19/, L.I.Timofeyev /15/, B.V.Tomashevsky and U.N.Tynyanov /16/ had a great influence on later works on rhythm. Even in the study of meter a successful attempt has been made to correlate meter and meaning (a range of images and themes) /10/.

There is another trend in the investigation of rhythm which is not widely acknowledged, but which appears to be very promising, being connected with other sciences of Man and capable of opening up avenues to the study of verbal and non-verbal thinking. An attempt has been made to see rhythm "from the inside" through the unity of a poetic image and the overlapping of semantic fields. Rhythm is considered as an intermediate stage between the continuity of Thought and the descreteness of Language. Rhythm is perceived subconsciously and is directed straight to continuous image thinking /11/.

Thus, a further perspective in the study of rhythm lies in a systematic approach to this problem, in the comparative study of the rhythm of different texts, different languages and groups of languages, and in the study of both verbal and non-verbal rhythms.

An extremely fruitful and valueable, if complex, approach would result from considering the concept of rhythm with reference to Man as the central object of investigation.

REFERENCES

/1/ P.Fraisse, "Les Structures Rhythmiques", Studia Phychologica, Louvain Publ. Univ., 1956 , 124 p.

/2/ M.Leberemam and A.Prince, "On Stress and Linguistic Rhythm", Linguistic Inquiry, 1977, 8, 2, p.336-449.

/3/ G.Lehiste, "Rhythmic Units and Syntactic Units in Production and Perception", Journal of the Acoustic Society of America, 1973, 54, 5, p.1228-1234.

/4/ P.Kiparsky, "The Rhythmic Structure of English Verse", Linguistic Inquiry, 1977, 8, p.189-347.

/5/ J.D.O'Connor, "The Perception of Time Intervals", Progress Report, London, University College, Phonetics Laboratory, 1965, p.11-15.

/6/ J.D.O'Connor, "The Duration of the Foot in Relation to the Number of Component Sound-Segments", Progress Report, London, University College, Phonetics Laboratory, 1968, p.1-6.

/7/ H.Woodrow, "A Quantitative Study of Rhythm", Archires of Psychology, 1909, 14.

/8/ А.М.Антипова, "Ритмическая система английского языка", М., Высшая школа, I984, II9 с.

/9/ Е.А.Бурая, "Роль просодии в формировании ритма спонтанной диалогической речи (на материале английского языка)", сб.науч.трудов, М., МГПИИЯ им.М.Тореза, I982, вып.I96, с.I0-30.

/I0/ М.Л.Гаспаров, "Семантический ореол метра. К семантике русского трехстопного ямба", в сб. Лингвистика и поэтика, М., I979, с.282-307.

/II/ Ж.А.Дрогалина, В.В.Налимов, "Семантика ритма: ритм как непосредственное вхождение в континуальный поток образов", в сб. Бессознательное, Тбилиси, I978, Мецниереба, т.3, с.293-300.

/I2/ Д.К.Исхаков, "Особенности семантической связности текстов английских лирических стихотворений ХУI, XIX и XX веков", АКД, М., I983, 2I с.

/I3/ А.М.Пешковский, "Стихи и проза с лингвистической точки зрения", сборник статей, Л.-М., I925, с.I53-I66.

/I4/ "Ритм, пространство и время в литературе и искусстве", Л., I974.

/I5/ Л.И.Тимофеев, "Теория стиха", М., Госполитиздат, I939, 232 с.

/I6/ Б.В.Томашевский, "Стих и язык", М., Госполитиздат, I959, 470 с.

/I7/ Н.Б.Цибуля, "Роль интонации в структурировании текста", АКД, I982, 24 с.

/I8/ Н.В.Черемисина, "Ритм и интонация русской художественной речи", АДД, М., I97I, 48 с.

/I9/ Б.М.Эйхенбаум, "Мелодика русского лирического стиха", в кн. О поэзии, Л., I969, с.327-5II.

PHONETIC AND PHONOLOGICAL COMPONENTS OF LANGUAGE RHYTHM

REBECCA M. DAUER

ESL Program
University of Massachusetts
Amherst, MA 01003 U. S. A.

ABSTRACT

Rhythm, or the grouping of elements into larger units, is a property of all languages. The particular rhythm of a language is the result of the interaction of a number of components, including phonetic components, such as the relative length, pitch, and segmental quality of accented and unaccented syllables, and phonological components, such as syllable structure and the function of accent. A system of rating whereby these components are broken down into features which can be assigned a plus or minus value allows us to compare the rhythm of languages or language varieties. Languages which have "strong stress" or which have been labeled "stress-timed" are seen to share certain features. Rhythm is a total effect involving phonetic and phonological as well as segmental and prosodic phenomena.

INTRODUCTION

Is it possible to develop a phonetic concept of rhythm that can be applied to all languages, in the same way that we use the system of cardinal vowels or the IPA chart of consonants? The distinction between stress-timed languages and syllable-timed languages [1, 2] is just such an effort at a general phonetic definition. In this theory, stress-timed languages show a tendency for stresses to recur at regular time intervals, and in syllable-timed languages, syllables are said to recur at regular intervals; all languages are believed to have one or the other rhythmic basis. Although many linguists have adopted the distinction, some have criticized the theory for being too simplistic (after all, it only divides all the languages in the world in half) and for grouping together languages which are felt to have noticeably different rhythms, such as English and Arabic [3] or Spanish and French. In addition, many native speakers of "syllable-timed" languages have objected to the designation, as if it somehow meant that their language had no rhythm. Indeed, Crystal and Quirk [4] refer to the lack of regular stress-timed pulses as "arhythmic." Linguists have had difficulty applying the concept to languages. Attempts to do so by instrumental analysis have been futile. Numerous experiments have shown that a language can not be assigned to one or the other category on the basis of instrumental measurements of interstress intervals or syllable durations [5, 6]. Moreover, Scott, Isard, and Boysson-Bardies [7] have shown that the perceptual tendency towards isochrony of stress beats is not specific to stress-timed languages, nor to language. Miller [8] had English and French phoneticians and nonphoneticians listen to selections of seven different languages and evaluate them as stress-timed or syllable-timed. Only Arabic was unequivocally categorized as stress-timed by all groups. Phoneticians generally agreed in finding Spanish stress-timed and Yoruba syllable-timed, but found no strong tendency for Finnish, Japanese, or Indonesian, and disagreed about Polish. This experiment seems to suggest that a language may be categorized on the basis of how strong and easily perceivable stress is.

Should we then give up the only phonetic theory of rhythm that we have, or perhaps turn to a purely phonological approach? Phonologists in the tradition of Trubetzkoy have treated rhythm in terms of the function and location of accent in the word. Metrical phonologists (following Martin [9]) have assumed that all languages have an underlying strong-weak distinction and show a tendency towards alternation which can be shown in a grid or tree structure of the word. Although this approach brings out the importance of grouping of elements into larger units, which is considered essential in all psychological definitions of rhythm, it tends to make all languages look alike, at least on paper, and makes no attempt to specify further how these patterns are realized in spoken language in continuous speech. But as Ladefoged and Wu [10] have noted, phonetic details are part of linguistics and do matter to any linguist who wants to make a complete, accurate description of a language.

It seems that an adequate description of rhythm in a language or across languages requires both phonetic and phonological information (a conclusion also reached by Hyman [11]). We can define rhythm as the grouping of elements into larger units; the units need to have some similarity and be marked off from each other in some way in order to be perceived as groups [12, 13]. In language, most would agree that the elements that are grouped are syllables, and that in some languages at least, stresses (or accents) serve to set off groups. Neither "syllable" nor "stress" have general phonetic definitions, which from the start makes a purely phonetic definition of language rhythm impossible. All instrumental studies as well as all phonological studies have had to decide in advance where the stresses (if any) fall and what a syllable is in the language under investigation in order to proceed. Although rules for syllable division and inventories of syllable types have been worked out for many languages on the basis of phonological criteria, stress is more problematic, and definitions of it have varied widely. In this paper, I shall use the term "accent" as it has been defined by Trubetzkoy [14] as the phonological feature which when realized promotes the perception of one particular syllable (or mora) in relation to others. Accent can then serve as a basis of rhythmic grouping. The term "stress" will be reserved for the phonetic realization of certain kinds of linguistic accent. I hypothesize that all languages have rhythmic grouping, but that not all necessarily have accent. Rhythm is a total effect (also probably a grouped series of motor commands in production) that involves the interaction of a number of components, of which the following appear to be the most important for the purposes of comparing languages. It is most evident in continuous speech through the repetition of rhythmic groups at a natural speed for the speaker. Obviously, some speakers and some styles exhibit better rhythm than others and seem to be more representative of a particular speech community. The following analysis is based on the "consultative" style [15] of informal lectures, monologues, or prose reading by people who are used to reading aloud. In each category, a plus, zero, or minus is assigned to a language depending on the extent to which it exhibits the feature in question.

COMPONENTS OF LANGUAGE RHYTHM

1. Length

Duration

\+ Accented syllables, and especially accented vowels, are regularly longer than unaccented syllables (by 1.5 or more). (e.g. English, Serbo-Croatian)

0 Accented syllables are slightly longer than unaccented syllables. (e.g. Spanish, Greek)

\- Accent does not affect the length of syllables, or the language has no accent. (e.g. Japanese, Yoruba)

Syllable Structure

\+ The language has a variety of syllable types (both heavy and light syllables with many different possible syllable structures), and heavy syllables tend to be accented, whereas light syllables tend to be unaccented. (English, Arabic)

\- There are a very limited number of syllable types (predominantly CV or CVC), and accent and syllable weight are independent. There may be active processes such as final cluster simplification, epenthesis, or liaison to break up or prevent the formation of unusually heavy syllables. (Spanish, French)

Quantity

\+ Quantity distinctions, if present in the language, are only permitted in accented syllables; in unaccented syllables they are neutralized (only short). (some Arabic dialects)

0 All quantity distinctions occur in accented syllables, but only a small subset can occur in unaccented syllables. (Estonian)

\- Quantity distinctions are permitted in both accented and unaccented syllables. Restrictions on quantity are not conditioned by accent. (Hungarian, Finnish)

2. Pitch

Intonation

\+ Accented syllables are turning points in the intonation contour. Pitch (usually high or changing) correlates with accent, but the actual pitch contour depends on the position in the utterance and the intonational meaning. Emphasis or contrast affects primarily the accented syllable. (English, Greek)

\- Intonation and accent are independent; there may be a negative correlation of pitch and accent. Relative pitch patterns may be consistent with respect to the word regardless of its position in the utterance or intonational meaning. Emphasis may affect unaccented syllables or be achieved by other means. (French, Japanese)

Tone

\+ Tones, if present in the language, only exist on accented syllables; unaccented syllables are atonal. (Swedish)

0 Tones are fully developed on accented syllables, but they are neutralized or subject to numerous changes (sandhi rules) in unaccented syllables. (Thai)

\- Tones are present on all syllables or all syllables with a particular structure, regardless of accent. If there are sandhi rules, they are not related to accent. (Yoruba)

3. Quality

Vowels

\+ The maximal vowel system exists in accented syllables; vowels in unaccented syllables tend to be reduced or centralized (especially open vowels). (English, Swedish)

0 The unaccented vowel system is smaller than that of accented vowels, but unaccented vowels are not necessarily centralized. There may be processes of devoicing or raising which occur only to unaccented vowels. (Russian, Portuguese)

\- There is the same vowel system and similar articulation in all syllables. If elision or devoicing processes exist, they affect accented and unaccented vowels equally and are determined by phonetic environment rather than accent. (Spanish, Japanese)

Consonants

\+ Consonants are more precisely articulated in accented syllables, and some may have special reduced allophones (e.g. syllabic consonants, loss of aspiration) or be subject to neutralizations in unaccented syllables. (English, Thai)

\- All consonants have the same articulation regardless of accent. Consonantal allophones are not conditioned by accent. (French)

4. Function of accent

\+ Accent can occur in different positions in a word (accent is "free" or free over a range) and is an integral part of the word shape for recognition. Moving the accent could result in a new word with a different meaning. (English, Spanish, Swedish, Russian)

0 Accent can occur only in one position in a word (accent is "fixed," typically on the first syllable). Moving the accent or adding an accent could result in the formation of a new word boundary. (Hungarian)

\- There is no word-level phonological accent; no one syllable consistently stands out over others in a word. Accent can be moved for stylistic or emotional reasons (in a language with a phrasal accent), but moving the accent does not result in a change in referential meaning or the establishment of new word boundaries. (Yoruba, French)

CONCLUSION

By applying these categories to various languages, one should be able to come up with a comparative rhythm "score." The more pluses a language has, the more likely we are to say that the language has "strong stress" ("dynamic" or "expiratory" accent) and is "stress-timed." The differences between accented and unaccented syllables are maximized, and accent would clearly be the principle for grouping. We would expect that naive native speakers--as well as trained non-native speakers--could fairly consistently identify accented syllables in continuous speech. In a language with many minuses in these categories, we would have to look elsewhere for the principle of grouping: what is it that permeates the entire linguistic system, binds units together and helps listeners segment the flow of speech into meaningful chunks? It could be patterns of tone, of syllable or vowel length, or even the repetition of certain segmental or grammatical features. Although the language may have some kind of accent, naive native speakers would have difficulty identifying the place of accent consistently in continuous speech, and linguists would have difficulty finding its acoustic correlates, even in words said in isolation. This does not necessarily mean that this kind of language is somehow defective or arhythmic because it is lacking a feature that certain prestige languages have. All languages have rhythm, but more independent research needs to be done to discover exactly what the rhythmic principles are in languages which do not show a tendency towards "stress-timing."

The above chart would also be useful in comparing different styles, dialects, or historical stages of a language. Rhythm can be significantly changed, for example, by pronouncing every syllable distinctly in a language which has vowel reduction rules (Jamaican English). Non-native speakers of English can improve their rhythm enormously by reducing unstressed syllables, and this is usually more successful than trying to get them to equalize stress beats. In comparing the naturalness of synthetic speech samples, Carlson, Granström, and Klatt [16] concluded that "the amount of isochrony implemented in the rules via, e.g., cluster shortening and unstressed segment shortening is probably sufficient, and no 'isochrony rule' per se need to be added." We must not forget that the division into segmental and prosodic phenomena is an abstraction created by linguistic science for the purposes of analysis. In early stages of language acquisition, Crystal [17] aptly notes that primitive words are used as units "with the segmental and non-segmental characteristics 'fused'." Even in adult language, segmental and non-segmental phenomena are interdependent and can influence one another. This fact is quite evident in the analysis of tone languages. It is also important in helping us to better understand language rhythm.

REFERENCES

[1] Pike, K. L. The Intonation of American English. Ann Arbor, University of Michigan Press, 1946.

[2] Abercrombie, D. Elements of General Phonetics. Edinburgh, Edinburgh University Press, 1967.

[3] Mitchell, T. F. Review of Abercrombie 1967. Journal of Linguistics, 5:153-164, 1969.

[4] Crystal, D. Prosodic systems and language acquisition, in Prosodic Feature Analysis. Edited by P. Léon, G. Faure, & A. Rigault. Montreal, Didier, 1970, p 77-90.

[5] Roach, P. On the distinction between 'stress-timed' and 'syllable-timed' languages, in Linguistic Controversies. Edited by D. Crystal. London, Edward Arnold, 1982, p 73-79.

[6] Dauer, R. M. Stress-timing and syllable-timing reanalyzed. Journal of Phonetics, 11:51-62, 1983.

[7] Scott, D., Isard, S., & Boysson-Bardies, B. Perceptual isochrony in English and in French. Journal of Phonetics, 13:155-162, 1985.

[8] Miller, M. On the perception of rhythm. Journal of Phonetics, 12:75-83, 1984.

[9] Martin, J. G. Rhythmic (hierarchical) versus serial structure in speech and other behavior. Psychological Review, 79:487-509, 1972.

[10] Ladefoged, P. & Wu, Z. Places of articulation: an investigation of Pekingese fricatives and affricates. Journal of Phonetics, 12:267-278, 1984.

[11] Hyman, L. M. On the nature of linguistic stress, in Studies in Stress and Accent. Edited by L. M. Hyman. Los Angeles, Dept. of Linguistics, Univ. of Southern California, 1977, p 37-82.

[12] Woodrow, H.. Time perception, in Handbook of Experimental Psychology. Edited by S. S. Stevens. New York, Wiley, 1951, p 1224-1236.

[13] Fraisse, P. The Psychology of Time. London, Eyre & Spottiswoode, 1964.

[14] Trubetzkoy, N. S. Introduction to the Principles of Phonological Descriptions. Edited by H. Bluhme. The Hague, Martinus Nijhoff, 1968. First German edition, 1935.

[15] Joos, M. The Five Clocks. Bloomington, Indiana University Research Center in Anthropology, Folklore, and Linguistics, 1962.

[16] Carlson, R., Granström, B., & Klatt, D. Some notes on the perception of temporal patterns in speech. Proceedings of the Ninth International Congress of Phonetic Sciences, vol. 2. Copenhagen, Institute of Phonetics, University of Copenhagen, 1979, p 260-267.

[17] Crystal, D. & Quirk, R. Systems of Prosodic and Paralinguistic Features in English. The Hague, Mouton, 1964.

METRO-RHYTHMIC AND PHONIC STRUCTURES OF SPANISH POETIC SPEECH

SERGEI F.GONCHARENKO

Translator's Department
Maurice Thorez Moscow State Institute of Foreign Languages
Moscow , USSR 119034

ABSTRACT

Spanish poetic speech has three systems of metro-rhythmic structures (syllabo-tonic, accentual-syllabic and tonic) and three varieties of phonic structures (ornamental, symbolic and paronymic). The metro-rhythmic and phonic structures play a major role in ensuring poetic communication through the means of Spanish poetic speech.

The mechanism of Spanish poetic speech has specific media of transforming discrete units of practical linear discourse into the continuum of what is known as score-spatial poetic signs /1/. The high communicative-informational potential of these "spatial signs" is largely determined by that specific impact which, in accordance with the author's communicative design, is exercised on the initial informativeness of speech units by the conditions of the poetic context, which contributes to combining all the informative elements of the poem "vertically" and "horizontally" into a cohesive communicative-informational complex - a lyrical text.

In other words, under the impact of the poetic context the informational relevance of units of poetic speech not simply becomes transformed (relative to their "homonyms" in the language at large) and not simply becomes multiplied, but also acquires a qualitatively new - aggregative-integral - character: remaining formally and in terms of factual semantics a combination of discrete linear units, the work of poetic speech as regards conceptual-semantic and aesthetic informativeness, is already found to be not the sum and in general not an arithmetical "product" of the semantics of discrete linear signs, but an aggregate and, to a certain extent, indivisible poetic macrosign none of what would appear to be its quite autonomous components can independently perform the communicative-informational function which it performs as part of a cohesive macrosign. As soon as a certain communicative element of the poetic play is taken out of the text, which severs its ties of similarity and/or contrast with other (and very frequently with all) elements of the poem, this element changes from a component of the spatial poetic macrosign (or perhaps from a spatial poetic microsign) into a flat linear - prosaic and discrete - sign.

Here, however, it is important to note another - on the face of it, paradoxical - property of connected poetic speech. Alongside an expressed tendency toward the loss by elements of the spatial poetic microsign of their autonomous communicative value, the opposite tendency is at work - a tendency toward the orientation of the conceptual structure of the units of the poetic discourse to the general conceptual and aesthetic informational programme of the text: each of such units, down to an individual word, being "overturned into the subject and idea of the artistic design" /2/, tends, as much as possible, to reflect the overall communicative task of the text in an integral, if simplified, form. This dialectical unity of the synsemantics and autosemantics of spatial and poetic microsigns predetermines the specific complexity both of the analysis of the poetic text (especially written in a foreign language) and of its poetic translation, i.e., of the synthesis of a spatial poetic macrosign equivalent in terms of conceptual-aesthetic informativeness in the target language.

An important role in the "spatialisation" of poetic speech units is played by poetic (metro-rhythmic, phonic and metalogical) structures, quite specific to the poetry of each national language.

Metro-rhythmic structures can be regarded as techniques of the systems segmentation of speech into verse lines as well as of the systems intralinear speech organisation which takes the form of an ordered alternation of marked and unmarked syllabic positions. The ideal pattern of such alternations is traditionally known as metre. The realisation of the metre in empirical verse is known as its rhythm. Implementing the principle of repeat at the syllabic and linear level, the metro-rhythmic structures ensure the generation of verse speech and form the basis for its development into poetic speech, in other words, speech capable of performing the function of poetic communication - communication with the laconic means of two-tiered semantic (factual and conceptual) and multiaspect aesthetic (aesthetic proper, cathartic, hedonistic, axiological, suggestive-hypnotic, structural-formal, functional-formal, etc.) information.

The metro-rhythmic structures play the major communicative role of "stratifiers" of semantic information: "linear" syntagmatic connections, which take no account of the "desyntagmatisation" (the term of I.R.Galperin) of verse speech, ensure a continuum of factual information, which bears a marginal character in the lyrical text and, as a rule, is of no basic significance. The "vertical" (naturally, combined with "horizontal") syntagmatic and paradigmatic connections determined by the metro-rhythmic division of the text guarantee the continuum of conceptual and aesthetic information.

Contrary to a widespread view, the Spanish metro-rhythmic structures correspond not to one, but to three different versification systems. First, to the syllabo-tonic system, whose metric repertoire in principle is as diverse as that of Russian or English syllabo-tonic poetry, including bisyllabic and trisyllabic metres. Even among the best-known poetic works by Spanish and Latin American authors the body of "flawless" syllabo-tonic texts, according to preliminary calculations, comprises no less than 3,000 poetic lines. Second, they correspond to the accentual-syllabic system, which represents a non-footed (in contrast to the syllabo-tonic system) compromise between the tonic and the syllabic systems. The accentual-syllabic metre in principle coincides with the metric pattern of dol'nik but, in contradistinction to it, presupposes the constant isosyllabism of the verse lines. Therefore the metre of the accentual-syllabic verse is determined both by the number of ictuses and by the number of syllables, for instance, three-ictus octosyllable, two-ictus pentasyllable, etc. Finally, old Spanish poetry and its modern-time derivations show a trend toward

the tonic system proper - toward what is known as taktovik ("Cantar de mío Cid") and dol'nik (old romances, some texts by Pablo Neruda and García Lorca, etc.).

The phonic verse structure can be interpreted as a device of the systems use of the grapho-phonemic repeats designed to convey semantic (as a rule, conceptual) and/or aesthetic information in a poetic text. The grapho-phonemic system of Spanish occasional alliteration contains 21 units (Spanish has 25 phonemes and 30 letters) and the grapho-phonemic system of the Spanish rhyme has 25 units, which, however, are not fully coincidental with the units of the Spanish phonological system.

The phonic structures, which, in the main, convey only aesthetic information, will be referred to as ornamental. But it should be observed that ornamental phonic structures may also exhibit a measure of semantisation, but not to the extent of enabling the grapho-phonemic repeat to gain the status of a quasi-morpheme /3/ - a sound-letter combination having a certain occasional meaning within a euphonic context, as, for instance, in the following verse by Miguel de Unanumo: "Esta es mi España" (the quasi-morpheme "es" meaning "existence" or even "eternal existence" or "Las montañas de mi tierra/ en el mar se miran" (the "distant" quasi-morpheme "mnr", whose meaning, "mountain-sea", symbolizes Biscaya as a land of mountains and the sea).

The phonic structures which contain a quasi-morpheme and therefore are undoubtedly semantized can be classed with symbolic phonic structures. Finally, clearly semantized ("quasi-morphemic") sound-letter repeats perceived as a specific phonic feature not so much of the verse line as of concrete lexemes /4/ which enter in image-paronymic relations can be categorized as paronymic phonic structures: manzana amanecida (Jiménez), avienta tus destinos al viento aventurero (Greiff), amarillas mariposas. (Jiménez).

In our time the recognition of the conceptual relevance of symbolic and paronymic phonic structures appears to encounter less and less resistance - at any rate, among the authors of works on linguopoetics and linguostylistics. However, as before, debates as to whether it is correct to speak of any informativeness of that part of the phonic structures which has been categorized as their ornamental variety continue unabated. Indeed, what, for instance, is the informational load of the sound repeats l and ll (i.e., of the repeats of the sound-letter "L", which in the Spanish grapho-phonematic system of occasional sound repeats unites both phonemes and both graphemes) in the following lines by Juan Ramon Jiménez (the general language probability of the frequency of occurrence of this graphophoneme is 5.7 per cent, and in this fragment its probable expectancy is exceeded more than twice, equaling 12 per cent)? " ... levanta nubes de polvo/ y llora con sus esquilas, /bajo la luna de oro./ La aldea del valle está/ quieta en humo blanco. Todo/ lo que era alegre al sol, sueña/ no sé qué amores lloroses ... "

The functional relevance of precisely such and similar phonic structures, more often than not, raises doubts. Some researchers simply deny that these phonic structures have any conceptuality of their own, reducing their communicative load to the creation of a sonic pattern, emotional resonance, etc.

It appears, however, that the inclusion in the conceptual-terminological apparatus of poetics of such notions as semantic and aesthetic information makes it possible to minimize misunderstandings and narrow the scope of theoretical confrontations of minor importance on this question as well.

In the author's opinion, ornamental alliteration (i.e., the phonic structures which carry no direct semantic load) bears an informative character if only because it is a major factor creating general textual information and imparting to the text the conclusive, coherent, integral character of the only adequate unit of poetic communication, of a "spatial" sign, which has a paradoxically indiscrete nature.

Furthermore, ornamental phonic structures are also discrete bearers of quanta of aesthetic and sometimes even mediated semantic information in its different manifestations, of which reference will be made to just a few.

1. First, it is a variety of hedonistic information connected with the identification of the degree of creative mastery and the technical freedom of versification and with the "delight" experienced by the recipient, who shares with the author "emotions of creative power over the language" /5/.

2. Second, it is the additional information that the recipient deals not with "practical", but with "poetic" speech, the adequate perception of which requires the use of a specific poetic code.

3. Third, it is phono-motivational information supporting by sound identities the semantic similarity or difference between the verbal structures involved in the sound repeat. Here two situations can be distinguished in their turn (A). Sound repeat as the phonic motivation of a trope construction, which emphasizes not only the associative-semantic, but also the phonic community of the comparatum and/or the tertium comparationis and/or comparant: "una lagrima lucero (Jiménez), "mi ilustre soledad de esquila y lana" (M.Hernandez) (B). Aid with phonic means to the reconstruction of lexical meanings in autological, i.e., formally ugly construction of a poetic text and the sound-letter motivation of the "fluctuating signs of meaning" (Yu.N.Tynyanov) in a poetic word /6/, as a result of which intensifies the semantic interpretation of units of syntagmatic sequences: "tremulos trigales" (Greiff), "flautas flexibles" (Hernandez), "valles llenos de dulce añoranza" (Jiménez).

4. Fourth, it is information indicating textual cohesion, expressed by means of phonic cohesion, which itself can ensure such a high degree of the coherence of a poetic discourse that it becomes possible to neglect the ramified grammatical and lexical devices of interphrasal connection.

5. Fifth, it is so called anagrammatical information, produced by the saturation of the text with grapho-phonemic complexes which form part of a certain key word.

6. Sixth, ornamental phonic structures are sometimes used for conveying onomatopoeic information. Naturally, in conceptual terms, it is a fairly lightened sound-letter device. However, its modest semantic advantages do not at all compromise this type of the phonic structure as an auxiliary means of sound expressiveness.

7. Seventh, the conceptualness of another variety of ornamental structures is quite possibly based on the latent effect of primary or secondary sound symbolism. The author says "quite possibly" since specialists in linguopoetics have not arrived at a final solution of this question, in distinction from, for instance, the problem of contextual sound symbolization of poetic speech. It is another matter that the "practical", primary "meaning" of a sound and its relevance to poetic speech, contrary to the conceptions of the majority of the authors of phonosemantic researches, are divided by as deep a chasm as the meaning of the word "dream" in everyday speech and in the text of Calderon's drama Life Is a Dream.

REFERENCES

/1/ N.I.Balashov, "Structural-Relational Differentiation of Linguistic Sign and Poetic Sign", Izvestia AN SSSR, Seriya literatury i yazyka, 1982, vol. 41, No 2, pp.125-135.
/2/ G.O.Vinokur, "Selected Works on the Russian Language", Moscow, 1959, p. 247 (in Russian).
/3/ V.P.Grigoryev, "Poetics of a Word", Moscow, 1979, pp.286-290 (in Russian).
/4/ Ibid., p. 272.
/5/ B.A.Larin "Aesthetics of a Word and the Language of a Writer", Leningrad, 1974, p.61 (in Russian).
/6/ Yu.I.Levin, "Concerning Some Features of the Content Plane in Poetic Texts", in Structural Typology of Languages, Moscow, 1966, pp.213-215 (in Russian).

HOW THE AUTHOR CONVEYS THE RHYTHM OF PROSE

Galina N. Ivanova-Lukyanova

МГПИИЯ им. М. Тореза
Москва 119034, Остоженка, 38

Grammatical form of text and punctuation marks will help define the rhythm of any written text.

However, many intonation variation are possible and they determine the exact rhythmic characteristics.

Thus the choice of a variant has to be made in keeping with the author's intention to change the rhythm of the text.

There is an opinion that the rhythm of work of prose can be studied on hearing a text, and that its written form doesn't give any information about its rhythmic and intonational structure.

However many investigators of rhythmic perculiarities of Russian belles-lettres texts didn't need to reproduce these texts orally. Therefore in the works of B.M.Tomashevsky[I], A.M.Peshkovsky[2], G.P.Firsov and others even the possibility of comparing the written and spoken forms of the same text isn't mentioned.

Thus an investigation of the written text of "Queen of Spades" gave B.Tomashevsky the opportunity to write on the rtythmic peculiarities of Pushkin's prose[I].A.M. Peshkovsky, while analysing the rhythmic structure of I.S.Turgenev's "Verses in prose" gave a beautiful specimen of a subtle pénetration into the rhythmic substance of the work.

Some linguists have asserted directly that the phonetic system of a piece of prose, which was created by the author, can be reproduced exactly while reading the text. For example, L.V.Scherba[4] wrote on the only right way of pronouncing a text, which corresponded to its correct interpretation; N.I.Zhinkin[5] considered that the reader must be able to find the very intonation which the author had in mind. Besides, it has been asserted experementally that the author, while choosing the very varient, which corresponds to the written form of the statement, reproduces in his inner speech all the variants, including the final one, which became the written form. Thus we may speak about both the phonetic similarity of the oral and written variants of the same text .

The author's intention as regards the rhythmic and intonational structure of the text is realized through linguistic devices which include the syntactical system of the text; and through non-language devices, which include, in their turn, a graphic arrangement of the material and punctuation marks.

The Rhythm of emotive prose is considered as the regularity of alternating: a) stressed and unstressed syllables, b) borders of sense division and c)Rising of combinations of rises and falls of tone gives us a chance to judge the level of its rhythmical harmony, which implements such things as smoothness, musicality of the text, and this is turn brings about a more suitable pronounciation of it. In belles-lettres the number of deviations is far less than in non-fiction. That's why its intonation is more versitile and flexibile. In belles-lettres intonation we find tone-contrasting intonational types, which seldom can be found in business like texts, journalistics and scientific texts.

The cause of such deviations may be found in the peculiarities of syntactical links of different functional texts. Subordinate construotions prevail in the texts of different functional styles (such as business-official, publicistic and scientific styles) in contrast to belles-lettres and colloquial texts in which subordinate, co-ordinating and asyndetic constructions are used equally. The high degree of intonational rhythm in a text, first of all, depends on the frequency of co-ordinating and asyndetic links which require the falling tone (pitch). The frequent combinations of the falling and rising tone (pitch) create and undulating movement of the intonation accepted as one of the elements of good rhythmic organization of the text.

The dependence of rythm on the syntactical structure of the text suggests that the author with the help of syntax determines the intonation of the text which is an inseparable part of prosaic rhythm.

It is also common knowledge that when we sound-track a written text the intonation may be different in variation. This is explained by the complex relation between syntax and intonation.

We tried to solve this matter by thinking thus: there are two types of intonation: the rising and the falling tones. The rising tone conveys the syntactical meaning of dependence and incompleteness, and the falling tone conveys the meaning of independence and completeness. If some syntactical structure has meanings that are combined in one intonational type then in this case it has no intonational variants (variations): if the syntactical structure conveys meanings that cannot be combined in one intonational type then it permits intonational variants, i.e. may be pronounced with both a falling and rasing tone. A non-terminal syntagma of a simple sentence, a terminal syntagma of declarative and interrogative sentence, and some others may serve as an example of non-intonational variants. A coordinating link inside a simple or a complex clause in which the meanings of completeness and incompleteness are not combined in one intonational type would be an example of a syntactical structure permitting intonational variants. The choice of this or that variant while reading must follow the requirements of the general message of the text which calls the author very often determines the choice of the intonation type using graphic means, when for example A.S.Pushkin uses any punctuation marks of his choice, i.e. the semicolon instead of a comma in unextended homogeneous sentences he points out necessity of the falling tone in cases when the syntax permits intonational variants.

Thus the rhythm of a written prosaic text may be suggested by the author's consent and in order to determine the rhythmic characteristic of a text it may be enough to use only its written variants.

and Falling of the pitch. Diversions from a regular pattern in these alternations comprise the main rhythmic characteristics of the text: syllabic, having sense-group character and intonational. If the syllabic characteristic doesn't permit any differences[7] in the written and spoken forms of the text, then the two other characteristics depend on the concrete interpretation of the text and cannot be derived from its written form.

The sense-group characteristic, which shows the regularity of the arrangement of sense-group limits depends on the quantity of contrasting sense-group differing in length (contrasting sense-groups are those, the length of which exceeds two phonetic words).

Thus this characteristic is based on the sense-group segmentation of the text which, as it is known, is variational. One reader divides the text into short sense-groups, another into long ones. It may seem that the rhythmical characteristic wholly depends upon the segmentation, cannot be more or less stable for a certain text. Still this is not quite so. Research testified that both written and spoken texts, from the rhythmical point of view, had a high degree of rhythmican composition of sense-group, which occurs irrespectively of any kind of reading. This means that the segmentation of the text into long or short sense-group leads to more or less similar results, when counting characteristics.

It is explained by the fact that the sense-group rhythmics is assessed not only by the absolute length os seens-group but also by a correlation of short and long sense-group, as the rhythmical side is violated by the neighbouring lenght-contradictory sence-group.

Thus the author's intentions as long as the rhythmical point of view is concerned provides for a certain constancy of this characteristic along with any correct spoken representation of the text. But this rhythmical representation is possible only due to correct reading, during which the reader understands in a correct way the whole rhythm of the text. That's why it would be extremely unnatural to read (quite) a rhythmical text dividing it atternately into short and long sence-groups.

A rhythmical tendency, which means our desire to equalise the rhythm of speech helps to assess the rhythmical side of the text. In spoken words this can be seen in a more rapid or slower pronunciation of different words and sense-group, in oral and written speech - in the equalisation of spoken passages, which happens due to the addition or interchangability of the words due to various stress-marks you place on unstressed words. Hence this rhythmical tendency determines the syllabic rhythmical characteristic.

Unstable stressing of link-words, pronouns, short numerals and adverbs is explained by the fact that these eradynamially unstable words serve as rhythmic organisators in speech. When they are found between two stressed syllables of the neighbouring words, they lose their personal stress, when found in a large interstressed interval, obtain it, thus showing the rhythmical harmony of speech.[8]

The third, intonational rhythmic characteristic depends upon the rhythmical tendency. An ideal rhythm in our understanding, a rhythm with interchangability on the upswing and descend in sense-groups. The correlation of identical intonational types is a deviation of the ideal rhythm. The author's rhythm is again a matter of determination by a syntactical order of the text. A certain character

Bibliography

I. B.V.Tomashevski. "About Verse", "Priboi", L., I929.
2. A.M.Peshkovski. "Rhythm in I.S.Turgenev's "Verses in Prose".
3. G.P. Firsov
4. L.V.Sherba. "**The Experience** of A.S. Pushkin's Linguistic Analysis of Verses", "Reminiscences".
5. N.I.Zhinkin. "The Development of Written Language by Pupils of the III-IV forms", "Izvestia", APN, RSISR, issue 78, p.I43.
6. A.N.Sokolov "Inner Speech and Thought" M., I968.
7. G.N.Ivanova-Lukyanova. "Rhythm in Prose", From the book "The Development of Phonetics in the Modern Russian Language", M., I973.
8. G.N.Ivanova-Lukjanova "About the Stress of the

IRREGULAR RHYTHMS

NIKOLAI KLIMOV

The Maurice Thorez Institute
of Foreign Languages
The Department of German Phonetics
Moscow Ostozhenka 38

Abstract

Speech rhythm is treated as an irregular temporal structure, characterized by a certain correlation of values of phonetic parameters of elements, comprising the structure itself. The problem of language specific feature of rhythm is discussed.

In phonetics rhythm is mainly understood as a certain harmony of elements in speech, manifesting itself in their isochronism and isomorphism. Phoneticians are apt to consider regularity the gist of rhythm.

For the versification and music theory, which mainly determines the phonetic studies of rhythm in prose, the idea of harmony is naturally of great importance, because of the emotional impact of regularity. The question arises, whether harmony should remain the decisive criterion in determining the object of the studies of rhythm in phonetics.

Let's proceed from the fact that in sciences, related to phonetics, there is a widely accepted point of view, according to which rhythm in the broad sense of the word is "a temporal structure of any perceived processes, comprised of accents, pauses, division into sections, their grouping, duration correlation, etc." /1/. This approach allows us to look upon the recurrence of elements as possible, but not obligatory. As E.Benvenist puts it, the word "rhythm", connected with equal intervals and recurrence, used to be one of the subtypes of a broader meanings /2/.

The idea of rhythm as a structure does not mean the appearance of a new conception, but the preservation of the old tradition, which continues to exist along with the idea of rhythm regularity, which according to E.Benvenist, appeared in the 5th century.

Rhythm as a structure means above all a dialectical unity of division and wholeness of movement as perceived by man.

It seems obvious that rhythm presupposes a physical sequence of elements, which are perceived as relatively independent. A speech signal, whose parameters may become changeable according to continuous linear function, is naturally devoid of any rhythm. On the other hand, movement aquires a certain rhythm if its elements are perceived by man as having certain temporal relations. Phychologists, dealling with musical perception, believe that a physical pause of more than 6 sec. destroys the temporal correlation of elements and thus distorts the feeling of rhythm /3/.

Rhythm as a structure is also used in phonetics. It manifests itself in all kinds of description of accentual-syllabic structures of speech. The description of accentual-syllabic structures, traditionally accepted in phonetics, uses a limited number of qualitative features, such as the number of syllables, accent and the recurrence of stressed and unstressed syllables in the structure. Such an abstract model of rhythm, which (with certain restrictions) can be called metrical, can easily be observed and does not require a more detailed experimental analysis. The correlations of elements in the accentual-syllabic structure, reflected by the metrical model of rhythm, correspond to the scale of order and convey the most general relations of the "more - less earlier - later" type. The elements themselves (syllables) are represented without disclosing their inner prosodic structure. Finally, the metrical model of rhythm is static, it does not convey the dynamics of syllable transitions. Its only dynamic characteristic is the direction of changes in rhythm-creating parameters, which makes it possible to discriminate between rising, falling and other types of rhythms.

Making the metrical model more concrete by means of rhythm-creating parameters (intensity and duration both in the sillable and inter-syllable relations) would be of considerable interest.

In this connection it would be interesting to deal with the problem of rhythm creating parameters as the material substratum of speech rhythm. No doubt, rhythm, as any other structure, is to a certain degree independent of its substance. One and the same rhythm can also be transformed into different kinds of substances, e.g. the conversion of the rhythmical structure from non-sound into sound substratum (e.g. a hand movement and prosody) and also from one type of sound substance into another (the manifestation of rhythm in melody and intensity). On the other hand it would be wrong to think that rhythm in general is not connected with any kind of substratum. It is important not to mix the question of the possibility of substance conversion with the question of optimal relation of structures and substratum, which provides for functional reliability of rhythm. If it's correct that "all the speech elements can be relevant to rhythm" /4/, than it's also correct that one and the same rhythmical structure gains and loses in its definiteness, "transparancy", requires a kind of effort for its production and identification.

Rhythm is mainly related to the energy foundation of speech. The material substratum of rhythm is above all the respiratory system of speech organs. In the most explicit way this point of view has been expressed by R.Stetson and D.Abercrimbie. The structure of movement on the level of other subsystems of the speech organs (voice production and articulation) displays a certain parallelism with the structure of respiratory pulses, thus producing a delicate differentiation of rhythms. When melody and articulation "deviate" from respiration, the form-building function of rhythm is performed by the latter. Thus the perception of rhythmical variations is closely connected with intensity changes of speech signal, and not with the changes in fundamental frequency or voice quality.

Investigation of irregular rhythms makes it possible to introduce another essential parameter in the description of specific features of different languages. Correlation between language specific features and rhythm in general can hardly manifest itself on the level of isochronism. The latter is determined mainly by speech, aims and conditions of communication, style. The harmony of speech units in texts of one and the same language can fluctuate within a very wide range. Besides, any text reflects the specific features of a language.

Language specific features of rhythm depend on metric schemes that prevail in speech continuum. According to E.Sievers, in German and English trocheeic or dactilic rhythm prevails, while in Romance languages it is iambic or anapaest /5/. But the most essential features of speech rhythm lie in the sphere of its non-metrical features, i.e. such phonetic variations that create rhythmical differences in one and the same metrical structure.

Syllable dynamics seems to be one of the most essential non-metrical features of an accent group. Syllable dynamics is the distribution of values of velocity (acceleration) of changes in intensity within syllable impulses. From the point of view of psychology the character of syllable dynamics is determined by correlation between static and dynamic muscular tension of speech organs.

The problem of syllable dynamics has been more or less investigated in different branches of phonetic siences,

such as the investigation of syllable accents in Germanic languages (Danish jerk, gravis and acut in Swedish, "sharp" character of syllables in Rhenish dialects). Investigators point to double-peak/single peak syllables and the character of release as parameters of dynamic differences of syllable accents. The change of syllable dynamics as manifestation of the speaker's emotional state was experimentally investigated by F.Troyan, who used the musical terms "staccato" and "legato" to denote two polar forms of dynamics /6/. Syllable dynamics as a whole ballistic movement and as a phenomenon of speech norm remains up to now a problem, which hasn't been sufficiently investigated.

The character of distribution of values of prosodic parameters in syllables is another non-metrical feature of an accent group. In this respect we can evidently distinguish only two poles, which will serve to distribute these meanings: decentric and centric (contrastive) rhythm. Regarding centricity and decentricity as auditory images, we must point out that the difference between these auditory images will be based not only on complicated interaction of syllable duration, intensity and tone, but also on spectral characteristics of vowels in syllables. It can be assumed (with a great degree of certainty) that the degree of reduction of unstressed vowels will affect the perception of centric rhythm.

We assume that investigation of irregular rhythms is of great phono-didactic importance. In teaching phonetics practising regular rhythm pursues mainly aesthetic aims, while irregular rhythm creates a special language colouring, whose absence in the speech of a foreign learner (especially on advanced stages) is the main feature of a foreign accent. Irregular rhythms become even more important if we treat a language not only as a system of phonological oppositions, but also as a general pronunciation norm, which distinguishes one language from another. This approach deals with identity of phonetic phenomena and not with their differences, thus giving the investigator a chance to make general conclusions, which, in their turn, lead to the phonetic basis of the language (a phenomenon which is both well-known and uninvestigated). Irregular rhythms comprise one of the most important components of the phonetic basis of a language.

According to modern conceptions of speech physiology, the upper structure of speech movement is the final result of interaction between several hierarchically organized levels of controlling speech movement, each of them having its own units. The data obtained in the course of investigation of speech ontogenesis and aphasia give ground to assume that the hierarchical structure os speech movement includes a rather autonomous "deep" level of the production of rhythmical groups as whole units, which are characterized by both metrical and non-metrical features, conveying the specific character of each language. If it is so, teaching phonetics of a foreign language should include a special stage of practising specific features of rhythm. This stage should precede teaching intonation and sounds.

References

/1/ B.M.Teplov. Psikhologiya musykalnykh sposobnostey. M., 1947, p.269-270.

/2/ E.Benvenist. Obshaya lingvistika. M., 1974, p. 383.

/3/ E.Nasaykinskiy. O psikhologii musykalnogo vospriyatiya. M., 1972, p. 192.

/4/ D.Abercrombie. Vzglyad fonetista na strukturu stikha. /Novoye v lingvistike. Vol. IX, M., 1980, p.404/.

/5/ E.Sievers. Grundzüge der Phonetik. 4.Aufl. Leipzig, 1898, S. 217.

/6/ F.Trojan. Biophonetik. Wien, 1975.

THE STUDY OF RHYTHM IN RELATION TO METRICS

MARIJKE LOOTS

Faculty of Letters
Catholic University Brabant
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

ABSTRACT

It is argued that the majority of experiments into what is called rhytmical or metrical speech has no actual bearing on the study of metre. The question is raised whether it is at all useful to make metrical verse the object of experimental research.

INTRODUCTION

Rhythm

In the strictest sense of the word, *rhythm* is the perception of groups in a series of stimuli. Instead of being merely successive, the stimuli appear to be organized into groups on the basis of a difference in prominence between the stimuli. Rhythm is an everyday phenomenon for everyone who is in the happy possession of an old-fashioned clock instead of an electronic time indicator. The name 'subjective rhythm' is given to the phenomenon that in a series of equidistant stimuli that are acoustically identical, people hear the stimuli as grouped, with a prominent stimulus either beginning (falling or trochaic rhythm) or ending (rising or iambic rhythm) the group. The term 'objective rhythm' aplies to cases in which the stimuli are objectively different, either by nature or because the experimenter makes them so. Provided that the interval between the most prominent syllables does not exceed an upper or lower threshold, the events will appear to be grouped in the same way as in the case of subjective rhythm.

The recurrent theme in the study of rhythm is the question whether the perception of rhythm is the result of 'the arrangement of durable elements, or [...] the succession of more or less intense elements [...]' [1]. Any attempt at an analysis of rhythm has to steer a middle course between overemphasizing the temporal aspect or the succession per se and the difference in prominence between the elements. No matter which one is chosen, the notion of isochrony is central to the notion of rhythm.

Subjective rhythm hinges on the idea that some stimuli are perceived as prominent because they fall at isochronous intervals. Objective rhythm depends on the claim that intervals between stimuli that differ in prominence will appear isochronous if these stimuli follow one another in more or less regular alternation.

Being a perceptual phenomenon, rhythm looks back on a long history of experimental research. One only has to open a few back numbers of Perception and Psychophysics to see that rhythm is still a hot issue in the field of psychology.

Metre

Whereas the perception of *rhythm* is spontaneous, *metre* is recognized on a cognitive basis rather than perceived on a sensory basis. By using metre, poets can try to induce the sensation of rhythm in their audience. Whether they succeed depends on a great number of factors, the willingness of the audience to comply with the poets' intentions being one of them. Poets and audience should share a familiarity with certain metrical schemata and other literary devices responsible for structure, such as alliteration and rhyme. This shared knowledge makes it possible for poets to deviate from strict regularity: the aesthetic effect of metre exists by virtue of the tension between the strict pattern and its realization through the medium of language. The realization differs from the pattern, not only because the poet consciously inserts a deviation from the norm (e.g. a trochaic foot in an otherwise iambic line) but also because speech does not easily allow itself to be put in a straightjacket of strict regularity. Both concrete and abstract properties of speech will rebel against the notion of strict regularity. By the term 'concrete properties' we mean measurable parameters, such as, for example, the duration of speech segments. Here, strict regularity is made impossible by the fact that individual speech segments have an inherent duration. By 'abstract properties' we mean the syntactic and semantic organization of speech, which also asserts itself at the concrete level, for instance by affecting the values of individual speech segments.

Problem

In this paper the question is raised whether anything can be gained by applying the principles and methodology of rhythmical research to the study of metre. Three positions in the study of metre suggest themselves:

1) Metrical speech is subject to the same regularities as tone sequences and should be investigated in the same way; this approach denies speech its own character.
2) Metrical speech is subject to a tendency towards rhythmic structure in that the latter imposes a regularity on the speech material that is otherwise absent in the acoustic signal.
3) Metrical patterns manifest themselves in speech material by totally abstract means, which elude

any experimental approach.

In this paper it will be taken for granted that the first position is untenable. The article hopes to demonstrate that the second position will also prove unfruitful and that new research paradigms are merely begging the question. One seems to be forced to accept the third position, even if this may have unattractive implications for an empirical discipline.

THE INFLUENCE OF RHYTHMIC STRUCTURE ON SPEECH MATERIAL

The metrical foot as a prosodic unit

Let us assume for the sake of the argument that a metrical line of verse may give rise to the perception of rhythm.
In the line

sweet day, so cool, so calm, so bright

an iambic rhythm will be perceived. What seems to be primarily responsible for the division of the line into groups of two syllables is the linguistic structure, which may be called abstract in that it can be recognized in writing. Let us assume, however, that on top of that linguistic structure there is a contribution of concrete measurable parameters. This assumption is given support by the repeated claim that lines in iambic metre "sound different" from lines in trochaic metre. If this is true, metrical feet are not merely the conventions handed down to us by the Classical writers, but they may also have a perceptual reality.
At the outset of our investigation we hypothesized that the metrical foot behaved as a prosodic unit, very much like a phrase or a clause. As such, it would show the prefinal lenthening characteristic of these larger prosodic units [2]. This idea was tested in a number of experiments, two of which will be mentioned briefly.

We started with nonsense syllables [də da:] read either with a falling rhythm or with a rising rhythm. The hypothesis predicted that the stressed element in the iambic line (foot-final syllable) and the unstressed syllable in the trochaic line (also foot-final) would be longer than their foot-initial counterparts. In the experiment involving nonsense syllables we found some evidence for prefinal lengthening in metrical feet, in perception as well as in production. Although this effect was only found when we asked our reader to scan the nonsense syllables (i.e. read them slowly and with emphasis) and although only non-naive listeners were able to ignore the actual opening of the line (stressed versus unstressed) we did find evidence for the metrical foot as a prosodic unit.
In the case of meaningful material, the set-up necessitated manipulating lines in such a way that they could be read as iambic (with an unstressed word at the beginning) or trochaic, depending on the metrical context.
The line was inserted as the second in a four-line stanza, the first line containing two-syllable words corresponding to the metrical pattern: trochaic in a trochaic stanza, iambic in an iambic stanza. The following will serve as an example (rather than giving a translation I will give a paraphrase in which there is a coincidence between metrical structure and language comparable to the Dutch instance; it should be noticed that in the Dutch example the critical second line only differs in the addition of an unstressed word at the beginning):

trochaic	
dreigend door het duister	threatening through the twilight
sprak een stem tot Piet	*spoke a voice to Pete*
handen in de hoogte	better stop your talking
stilte of ik schiet	silence, or I'll shoot
iambic	
opeens bewoog een tak	at once the air resounds
er sprak een stem tot Piet	*a voice to Peter speaks*
ga heen en kijk niet om	obey the voice you hear
je weet dat ik dan schiet	behave or I shall shoot

Of each line the individual segment durations were compared in their iambic and trochaic reading. No systematic difference between foot-final elements and their foot-initial counterparts was found [3]. Presumably, in the meaningful material a possible temporal organization of the material in terms of metrical feet was overridden by other prosodic regularities, governed by the syntactic organization of the material.
The above example should be read as an instance of how the many interactions in speech may override a possible rhythmic structure. In this example, as in all other variations upon this theme, metrical feet coincide with word or phrase boundaries in one type of rhythm, but, as a logical consequence of the opposition between rising and falling rhythm, never with the other kind of rhythm. If we want to investigate the interaction between syntactic and morphological structure on the one hand and metre on the other, we are faced with the problem that, in the majority of cases, it is impossible to vary the one and keep the other constant. This renders it impossible to make syntactic and morphological structure the object of experimental investigations into metre.

Isochrony revisited

The claim that in metrical speech prominent syllables should be separated by isochronous intervals is considered unrealistic in view of the fact that at a suprasegmental level the duration of speech segments is affected by many more influences than metre. If the claim of strict isochrony is considered naive, would not a more sophisticated interpretation of the notion be feasible? Two lines of research into the rhythm of speech have been developed in which the notion of isochrony plays a central part. The question should be considered whether these approaches might prove fruitful for the study of metre.

Rhythm is predictability. Martin [4] argues that the rhythmic structure of speech enables listeners to generate expectancies concerning later events in real time on the basis of the perception of earlier

events. The notion of isochrony is inherent in Martin's model in that the hierarchical organization of accent patterns depends on the notion of relative timing. Over time, rhythm has become more or less synonymous with predictability. In a series of experiments the influence of a disruption in the temporal structure on the predictability of a certain phoneme has been investigated, e.g. [5]. Rather than suggesting an application of this experimental design to the study of metrical speech, which is not unthinkable, we should like to argue that this interpretation of rhythm is misguided. Although it is true that rhythm entails predictability, predictability does not necessarily entail rhythmicality. Predictability may be the result of so many factors other than temporal structure that the question whether speech is predictable seems to have little bearing on rhythm.

Isochrony and perceptual centers. If not between the onset of syllables, vowel or words, isochrony is claimed to exist between 'perceptual centers' [6]. Perceptual centers correspond to the locus of the 'stress beat'. It has been demonstrated that P-center aligned digits are perceived as more isochronous than randomly aligned digits. Buxton [7] even reports on an experiment with meaningful material in which subjects found it easier to tap the rhythm to P-center aligned words than to randomly aligned words. It would not be difficult to think of an experiment into metrical speech in which the claim of metrical speech being isochronous could be tested. However, the location of the P-center has been proved to be so indeterminate that one wonders at the ease with which Buxton seems to have constructed her material. P-centers seem to be dependent not only on the nature of the initial consonant but also on the duration of the medial vowel and final consonant of a CVC sequence ([8], [9]).

CONCLUSION

In our own work, we have been unable to demonstrate that rhythm imposes a regularity on speech material at a concrete level. Rhythmic structure was supposed to manifest itself in the form of metrical feet. In various recent publications, the term 'rhythm' has become little more than temporal structure per se, or mere predictability. If we cast our net so wide as to include these interpretations of the terms, there is no end to our investigations into metrical speech. If we adhere to the more orthodox meanings of the words rhythm and metre, however, we will have to conclude that, no matter how much their interaction will continue to interest us, its precise nature can only be speculated on, and cannot be investigated with the means we now have at our disposal.

REFERENCES:

[1] Fraisse, P., Rhythm and Tempo, in: Deutch, D., <<The Psychology of Music>>, Academic Press, 1982.

[2] Klatt, D.H., Vowel lengthening is syntactically determined in a connected discourse, <<Journal of Phonetics>>, 1975, 3, 129-140.

[3] Loots, M.E., <<Metrical Myths>>, Martinus Nijhoff, The Hague, 1980.

[4] Martin, J.G., Rhythmic (hierarchical) versus serial structure in speech and other behavior, <<Psychological Review>>, 1972, 79, 487-509.

[5] Shields, J.L., A. McHugh, and J.G. Martin, Reaction Time to phoneme targets as a function of rhythmic cues in continuous speech, <<Journal of Experimental Psychology>>, 1974, 102, 250-255.

[6] Morton, J., S. Marcus, and C. Frankish, Perceptual Centers (P-centers), <<Psychological Review>>, 1976, 83, 405-408.

[7] Buxton, H., Temporal Predictability in the Perception of English Speech, in: Cutler, A., and D.R. Ladd, <<Prosody: Models and Measurements>>, Springer Verlag, Berlin, 1983.

[8] Fowler, C.A., 'Perceptual Centers' in Speech Production and Perception, <<Perception and Psychophysics>>, 1979, 25, 5, 375-388.

[9] Fox, R.A., and I. Lehiste, The effect of vowel quality variations on stress-beat location, <<Journal of Phonetics>>, 15, 1-14, 1987.

This paper has greatly profited from discussions with Els den Os and, at a later stage, Prof.Dr. A. Cohen, both of the Department of Phonetics, Utrecht, The Netherlands.

THE EVOLUTION OF OLD GERMANIC METRICS: FROM THE SCOP TO THE SCALD

Olga Smirnitskaya

Dept. of Germanic Philology
Moscow State University
Moscow, USSR 119899

ABSTRACT

In the evolution of Old Germanic poetry the need for a metre as the external verse form with the universal range of application is supplied in two ways. The metre is either abstracted from the concrete prosodic structures of epic alliterative line giving rise to syllabotonics (in Late Old English poetry) or is originated as the discovery of form within the language (the formalization of prosodic word-structures in scaldic poetry).

I. The alliterative verse

I.I. In works on comparative metrics the Germanic alliterative verse (AV) is usually referred to as a free form of tonic (or accentual) verse. Thus, M.L.West defines an alliterative line as "a variable unit containing two stresses and as much else as the poet saw fit to put in" /I,p.I8I/. This definition is to a certain extent contradictory. In fact, sentence stress is the basic measure of alliterative long line, but the long line carries not two stresses, but four. As for the short line, or half-line, it is known since Sievers' "Die altgermanische Metrik" /2/, that it not only counts stresses (phonetic words), but also takes into account - at least in regular forms of AV - the prosodic syllable features within the word. In its schemes syllables are classified by quantity and also by stress. However, it follows from the alliterative design that in the long line the gradation of phrasal stress is taken into account as well.

The complexity of AV, the union in it of utmost freedom with "inscrutable and needless" distinctions has always impeded scholars. Innumerable attempts have been made to explain away facts discovered by Sievers and his followers, either by changing priorities from the linguistic arrangement of the line to its oral (musical) performance or by drawing a sharp border-line between the principal (metrical) schemes and their phonetic (rhythmical) realization. "Sievers faßte bloß die Realisierung ins Auge",- Kuryłowicz writes in this connection, - ohne zum Grundschema vorzudringen" /3,S.I40/.

But Sievers' special rules are rules indeed, that is they prescriptively distinguish between the metrical (those recurrent in verse) and non-metrical (not in use or occasional) lines (cf./4,p.I74/).

Thus, being an accentual system, AV is quite sensitive to the quantitative structure of words in stressed positions. The change in the word order in Old Norse Vsp.42.7. fagrrauðr hani would result in a non-metrical ('too light') line *hani fagrrauðr. But a minor emendation would reconcile it with the scheme: eða hani fagrrauðr (cf. Gðr.II.7. eða gull glóðrauðr).

It is only natural that the metrical relevance of secondary stress provokes main objections. To quote Kuryłowicz again, "Nach Sievers gehören sie zum metrischen Schema, während sie in Wirklichkeit bloß eine submetrische Rolle spielen, d.h. beim mündlichen Vortrag zwar berücksichtigt werden, aber metrisch ebenso wenig relevant sind wie die kombinatorischen Varianten der Sprachlaute für die phonologische Gestalt des Wortes"/3,S.I40/. Nevertheless, in spite of the theory, secondary stress is essential to the metre in such a line as Old English Bēo. 463. Sūð - Dena folc (E-type, in Sievers' notation), where it supports the four-element structure of the line, i.e. prevents the weak syllables from slurring. Thus, this line is not to be modified into *mihtigan folc (but cf. Bēo. I398. mihtigan drihtne). E. Sievers, with all his alleged 'empiricism' made a clear-cut distinction between the schemes with obligatory (i.e. metrical) and facultative (i.e.submetrical) secondary stresses.

Language selectivity of AV is never so persistent as in its 'rhythmic licences'. The extrametrical (i.e.additional to two scheme 'drops') anacrusis in BEOWULF will be a good example. The actual words in this position make it evident that in spite of what is expected of the accentual verse, AV distinguishes

between the so-called 'phonetic words' and words as prosodically structured lexical units. In the regular verse of the BEOWULF-poet anacrusis is reserved for unstressed morphemes (i.e. preverbs), while the unstressed auxiliary words (i.e. prepositions, conjunctions etc.) are avoided in this position. The pattern of Beo. 2I7. Gewāt þā ofer wǣgholm / winde gefȳsed is recurrent in at least 75 lines of the poem. Thomas Cable took notice of the fact, pointing out that the deviations in metre are based on the same material as the deviations in Germanic accentual word structure /5,p.35/. Hence, it can be easily understood why anacrusis has remained in fact a minor rhythmic licence in the Scandinavian version of AV, where it occurs only occasionally and only in loose forms of Fornyroislag. The matter is that in Old Norse there are practically no words with non-initial stress (verbs with prefixes). Or, to say it in another way, there exists no prosodic material where anacrusis could have been opposed to 'metrical' initial drops in B-,C-types (x´x´ or x´´x).

These facts taken into consideration, we can better size up the essential difference between regular forms of AV and its loose forms. The rhythmic tendencies of the latter are in regular forms split into the main rule and the alternative rule, that is the rule realized under definable linguistic conditions (cf. Keyser's approach in /6/).

If metre is defined as an invariant scheme abstracted from the prosodic structure of a concrete line, then such metre is not to be found in AV, however complicated and strict it might be. The development of AV leads not to its abstraction but to its further splitting and devision into variant metrical (main and alternative) schemes and, consequently, to its still closer union with the poetic language.

If metre is called the general law of verse, then the alliterative metrical system can be with good reason compared with the common law of ancient Teutons with its incidental detailing and casuistry. I believe that this mode of existence of Germanic verse (that is its not being abstracted from the word- and sentence-prosodic structures) is precisely the feature that makes it so interesting for the theory of metrics. It is an archaic feature of verse typologically implied by the very essence of epic authorship.

I.2. Theoretical studies of AV have always been verified by the question of how an ancient scop could cope with the system so astoundingly complicated as this. A. Heusler believed that Sievers' five types of the short line are to be regarded as a feature belonging not to AV as such but to the artificial style that had developed in 'Leseepos' as a result of secondary normalization /7,S.I30/. Now that owing to the discoveries of recent decades, more is known about the nature of epic authorship, we would rather say : formal complexity of AV is the result of its 'artlessness', that is of the fact that the scop was not aware of verse-form as such. It was not the form as a system of devices that he mastered but the formally organized - formulaic - language. Recreating and varying formulas the scop was at the same time recreating the verse with all its gradual transitions between the norm (canonized forms), usage and free variation.

It would not be a mistake to say that the whole theory of AV is imbedded in the formulaic theory as created by M. Parry and A. Lord. Parry's initial definition of the formula as "a group of words which is regularly employed under the same metrical conditions to express a given essential idea" /8,p.80/ and still more the illuminating statements of "The Singer of Tales" by A. Lord point to exactly the same type of relations between language and verse as was outlined above, except for the fact that the 'tendencies' of the folklore epic songs as distinguished from the medieval epos, have not jelled into 'special rules'.

But in spite of Lord's assumption that the singer "learns the meter ever in association with particular phrases, those expressing the most common and oft-repeated ideas of the traditional story" /9,p.32/, and that, consequently, "any study of formula must therefore properly begin with a consideration of metrics and music" /9,p.3I/, the nature of epic metrics remains in the background of his theory. On the same pages he labels the verse "a more or less rigid rhythmic pattern" the singer "has to pour his ideas into"/9,p.22/ which seems exactly the opposite to the assumption cited above. Francis P. Magoun and his pupils by whose untiring efforts hoards of Old English formulas were examined, never attempted any explanation of AV, though 'the five types', as one could have imagined, proved indispensable for the practical purposes of classifying formulas. In the heated discussion of the Sixties around the formulaic theory (see the review of Ann Ch. Watts /I0/) the problems of verse were altogether omitted from consideration. This indifference to the verse aspect of formulas can be accounted for by two reasons.

Firstly, the formulaic poetry was invariably viewed by the followers of Parry and Lord as the oral-formulaic poetry, formulas being regarded as the visible trace of oral composition, that is "the composition during oral performance".

The 'metrics', in its turn, was mainly thought of as the reflection of 'music', the reconstruction of which was believed to be the major aim of scholarly studies. Special importance was attached to the Anglo-Saxon harp (cf. the discussion in /II/). The difference between the literary situation in the twentieth century Yugoslavia ind the situation in medieval England was obscured in that reasoning, though this particular difference is responsible for the fact that the epic poetry was transferred to parchment without the assistance of a philologist with a tape-recorder. It has been shown ever since that formulas can not be used as a proof of the oral origin of a text in its extant form. They do not fall into disuse, to say the least, in texts definitely known as those created on parchment. Formulaic poetry - and the corresponding type of metrics - lived as long as 'the unconscious authorship' (M. Steblin-Kamenskij's term) prevailed in medieval written poetry. To quote Steblin-Kamenskij, "The movement from unconscious to conscious authorship, obviously, is the basic direction in which literature develops. It is a great simplification, though, to assume that the transition from oral to written literature coincides with the transition to conscious authorship"/I2, p.I30/.

Secondly, among various types of formulas and formulaic systems it was the semantic formulas (i.e. formulas serving to equip epic themes) that mainly attracted attention. It is clear, however, that the connection of verse with language is first and foremost realized on the level of more general patterns underlying semantic formulas, i.e. on the level of rhythmic-syntactic formulas. As a matter of fact, only the latter justify speaking of the 'totally formulaic style' as the general organizing principle of Old Germanic poetic language. As for semantic formulas, their share is largely dependant upon the genre of the text and other features of its poetic style.

Being rooted in 'unconscious authorship', AV can not break through the range of traditional ideas and values of epic poetry, an this puts an end to its existence.

I.3. There were several attempts to explain the collapse of AV (first of all in the English tradition) as the result of the changed structure of the language. Primary importance was attached to the changes in word prosodics (quantitative changes in Middle English word) , in word-stock (numerous French borrowings and the deterioration of poetic vocabulary) and also to the analytic tendencies in grammar. This approach (first applied to AV by Winfred P. Lehmann /I3/) seems nowadays too straightforward. No less radical prosodic transformations (as well as other linguistic changes) took place in the prewritten period, but while an unbroken poetic tradition and the continuity of poetic texts were preserved, they did not produce any catastrophes in the metrical system. The Old Norse "post-syncopal" Fornyrðislag had, obviously, little affinity to the Common Germanic long line which was current at the time of syllable autonomy. Still, both systems, which are thousand years apart, are diachronically and typologically (as far as verse - language relations are concerned) identical. Linguistic changes become destructive for the verse only when poetry spreads to the spheres of reality unconquered (and unconquerable) by tradition, and the verse comes into contact with new subjects and raw speech material. This process can be to some extent traced in Late Old English texts. Thus, although the author of the Late Old English poem DURHAM takes great pains to follow the classical samples of AV, his attempts are bound to fail:"like a boy riding a bicycle, once the traditional poet or singer began to think about what he was doing, he was liable to fall off" /I4,p.I76/. Although the syllable range varies in DURHAM within the same limits as in BEOWULF and the number of alliterative words per line is usually observed, the alliteration, marking accidental words, is ineffective, the metrical schemes of the short line are crushed, and the place of formulas is taken by disorderly speech material. The verse of DURHAM might be defined as "a variable unit containing two stresses and as much else as the poet saw fit to put in" (see I.I), but this is no longer alliterative verse.

At the same time it is quite symptomatic, that in this particular loose verse and as an attempt to compensate for its looseness, the "alternating rhythm becomes almost mandatory"/I3,p.I00/:

9. Is in ðere byri eac/ bearnum gecyðed
´ x ´ x ´ x ´ ´ x x ´ x

I4. Is ðer inne midd heom/Æðelwold biscop
´ x ´ x x ´ ´ x x ´ x

The same is more or less true of other Late Old English texts. The way is gradually paved for the adoption of Latin and French poetic metres and for the conception of syllabo-tonics.

Some centuries earlier and on an incomparably larger scale the process of abstracting metrics as an external form began in the Scandinavian tradition. The starting point for this process was (as in the case of Old English poetry) the spreading of poetry to new subjects (first and foremost the sphere of the actual present) and the development of individual authorship. But the scaldic metrics unlike the budding syllabo-tonic schemes

of English poetry appears as the result of abstracting the form within the prosodic structures of the language itself.

2. The scaldic verse

2.I. The scaldic verse is generally considered as a tightened form of the epic verse: "the scalds added external requirements to those they had inherited"/I3,p84/. The scaldic line is formed, from this point of view, by adding a fixed cadence (´ x) to the short line of the epic mould. The scald retains the accentual schemes of the line (Sievers' five types), but adds to them some innovative syllabic restrictions on the number of unstressed syllables and the quantitative rules: so, resolution is permitted in the initial positions of the line but avoided in the middle of the line, etc. The alliteration, in its turn, is subjected to some new formal restrictions and assisted by internal rhyme ('hending') of two types : aðalhending, or full rhyme, in the even lines of a vísa and skothending, or partial rhyme, in its odd lines.

It is clear, however, that this approach to the scaldic verse entails additional complications to the old question of how poets were able to cope with their technique. The rules of epic poetry, as we have seen, were not cumbersome for the scop : "he learnes the meter ever in association with particular phrases" /9,p.32/. But the scaldic poetry is demonstrably unformulaic. Entity of sense is not evolved by its lines. So, the three words of the following line by scald Sigvat Þorðarson all belong to three different sentences interwoven within the space of a helming: eirlaust - konungr - þeira. Peter G. Foote might be quite right when he suggests that "pairs of alliterative words and rhyming stems must have hung together as tags, perhaps not 'formulaic' in the strict sense of the word, but ready to spring in mind"/I5,p.I83/, but the technique of such composition, - if we consider the remarkable 'scaldic sensibility'(Carol Clover) in matters of authorship, - remains even more obscure with this suggestion.

What is still worse, there are numerous lines that do not lend themselves to the routine procedure of metrical analysis: some of the generally accepted criteria are ineffective in case of the scaldic verse (especially those based on the semantic values of words), others - insufficient (as eddic alliteration in lines with an additional stress). The accentuation of the simplest line of four words (kilir ristu haf Lista) becomes a problem for a scholar attached to the traditional approach. It is often assumed that the scalds sacrified some metrical rules for the sake of some other metrical rules, and that being inveterate 'formalists' they often actually violated form. Sometimes they went so far as to put auxiliary words in the position of key- alliteration to adjust the number of syllables and the framework of hending.

We prefer another solution to this problem. It is the contention of the present paper that eddic metrical schemes are not to be taken for granted in scaldic verse. The violation of the eddic structures was not compulsory but deliberate, not chaotic but systematic in scaldic versification. In his attitude towards the outworn treatment of the language in traditional poetry the scald resorted to a device that might be called 'alienation'(ostrannenie) after the Russian formal school. He experimented with the prosodic structures of words like a true 'structuralist' and in doing so reduced the structures of language to a few operable patterns. We are going to show now that scaldic metrics is not at all as complicated as it looks.

Old Icelandic literature (unlike Old Irish) completely passes over the question of how 'the young scalds' learned their trade. It is well-known to all those who attempted to penetrate 'Snorri's categories', that elucidations of the author of the YOUNGER EDDA are in fact mystifying, and the most important things are left unsaid. Suffice is to say that abounding in most exotic terminology, Snorri's famous treatise does not even possess a coherent term for alliteration.

2.2. The metrical units.

In all probability the scaldic line comes from the alliterative epic line, but the relations between the two metrical units underwent complete transformation in scaldic poetry.

The scaldic six-syllable line derives from the short epic line. But the short line is the ultimate unit of AV which means, first, that it can not be divided into any smaller segments, and, second, that it is subordinated to the long line in the same way as the prosodic structures of the words are subordinated in speech to phrasal rhythm. It is then the long line with its rhythmic integrity and its variable schemes establishing the semantic values of words which can be rightly called the principal unit of AV.

This subordination of metrical units is abolished in scaldic verse, whose metrical schemes are constituted by the prosodic structures of isolated words. As a matter of fact, the continuity of the phrasal rhythm can not be retained in the verse where the phrase is broken by unjustified enjambements and is interwoven with other phrases to make the text nearly inscrutable. Alliteration under these circumstances provides only a formal connection for lines with the general accentual pattern. In other words, the long line ceases to exist in scaldic poetry as an integral unit. It is transformed into a distich, or a constructive element in the composition of a vísa (cf. the Old Norse term fjórðungr, i.e. 'the fourth part of a vísa').

Thus, in scaldic poetry the short line is autonomous and serves as the principal metrical unit or the line proper (vísuorð).

The autonomization of the short line is most evidently manifested in its cadence. The scaldic cadence is obligatorily marked by a hending and is formed by an inseparable (whole) word. Marking the end of the line, the cadence is at the same time an element of the binary structure of the line. The remaining part of the line is, in its turn, segmented into two separate 'prosodic words'. The successive binary segmentation of the line corresponds, as we can see, to the successive binary segmentation of a vísa (vísa - helmingr - fjórðungr - vísuorð). At the same time, there appears a certain correlation between the metrical form and the separability of scaldic kennings.

In the composite structure of scaldic lines the traditional metrical types are subjected to considerable simplification and deformation.

2.3. The metrical types of scaldic verse (dróttkvætt). Thus, the scaldic line consists of three 'prosodic words' that may be termed in accordance with the succession of its segmentation as the finale (cadence), the mediale and the initiale. The boundary of the mediale is marked by the boundary of the penultimate notional word:

the main part of the line		cadence
austr se'k /	fjoll af //	flausta
initiale	mediale	finale

Each of these three segments of a scaldic line has essentially different metrical functions and rhythmic possibilities. The rigidity of the line increases from its beginning to its end. The finale, as we know, is the line's constanta. The mediale is chosen by the scald from all the available prosodic structures of the language (some structures, however, are dropped out or merged into one, see below). In this respect the mediale can be termed the line's alternanta. The structure of alternanta fully predicts the metrical treatment of the linguistic material in the initiale, which owing to its predictability allows considerable rhythmic variation (varianta). It is within this section of the line that quantitative substitutions (sometimes doubling the syllabic range of varianta), additional word-boundaries and cohesion are widely practised by the scald. However, the immediate prosodic prototype of the initiale is the prosodic structure of a two- or three-syllable word.

Three prosodic types of alternantas are distinguished; hence, the whole variety of epic metrical forms (five types with their variations) is reduced to only three unified patterns. The lines which can not be confined to these three types (specifically, lines with weak initial position) are not infrequent in the earliest scaldic poetry (the 9th - the first half of the 10th century). In other cases they develop as a secondary metrical device in the innovative efforts of individual scalds; cf. for instance VELLEKLA by Einarr Skálaglamm and some of the varieties of dróttkvætt encountered in HATTALYKILL and HÁTTATAL.

Type I, neutral, alternanta – x ('heavy'). This is the only type, whose alternanta admits of the inner (additional) word-boundary, on condition that the weak position in it is filled by the lightest of 'clitics'. Thus, the finite forms of the verb are avoided in this position, the varianta (initiale) of the same type or alternanta of type 3 being reserved for them. Examples: hnekðumk / heiðnir // rekkar; rýgr kvazk / inni // eiga; setit hef'k / opt við // betra. Type I accounts for 39 (Þórbjǫrn Hornklofi) up to 56 (Arnórr Þórðarson) per cent of lines in scaldic poetry of 10th-11th centuries and has a conspicuous preference for odd lines.

Two other types of alternantas are formed by a 'minimal' word ('light' alternanta). Both of them are functionally marked and prefer even lines. Type 2, alternanta ◡◡ (short two-syllabic word): allsvangr / gǫtur // langar; berr mik / Dǫnum // ferri; varð'k þeim / feginn // harðla. Type 3 with a one-syllable alternanta: hialdrgegnir / bú // þegna; vindblásit / skóf // Strinda; kilir ristu / haf // Lista.

The one-syllable alternanta in type 3 has some noteworthy quantitative restrictions. Thus, the structures with a long vowel in a closed syllable and/or consonant cluster (skóf, batt, lézk) are permitted for finite verbs but avoided for nouns and nominal forms. This restriction (known as 'the Craigie's rule') reflects the accentual disparity between the verb and the noun in AV, that is their belonging to different metrical ranks: the nouns, like 'hrings, hraustr' appear to be too 'heavy' for this alternanta.

The prosodic distinction of 'heavy' (type I) and 'light' (types 2,3) alternantas is manifested in their treatment of alliteration and rhyme. The 'heavy' alternanta as a general rule (more than 85% of lines among hǫfuð-skáld, i.e. 'head-scalds') is marked by rhyme and/or alliteration. The 'light' alternantas, on the

other hand, take no part in sound repetitions irrespective of the meaning of words in this position. The main weight of the line is correspondingly shifted in such lines to varianta (the initiale) crammed with heavy syllables and marked by both rhyme and alliteration. The natural prosodic strucures (as they are reflected in eddic schemes) are substituted in these lines for an artificial, forced rhythm, most vividly shown in type 2 (with bisyllabic compounds in the initiale). Such lines as hugsvinn / kona // innan would be interpreted in terms of 'the five-type system' as a heavy variety of A-line (⊥ ⊥ ⊥ x); that is just the way they are usually interpreted in literature on scaldic poetry. The structural function of rhyme is sure to be denied in this case /16, p. 35/. It is evident, on the other hand, that this artificial rhythm was used as a most effectual tool of bringing into prominence scaldic nonce words heavily burdened with consonant clusters and dissected by rhyme and alliteration in juxtapposition. Cf. some more examples from Sigvat's ERFIDRÁPA ÓLÁFS HELGA (1040): (type 2) sóknstríðs / firum // ríða; margdýrr / konungr // varga; stálgustr / ofan // þustu; (type 3) úthlaupum / gram // kaupask; hundmǫrgum / lét // grundar: hjaldrmóðum / gram // bróðir; framlundar / Óg // mundar.

The term 'stress' has been avoided above, although the observed features of scaldic metrics can be easily described as 'taking the stress off' alternanta, 'the stress junction' in varianta etc. However such a description would obviously be simply re-encoding the facts, following from the quantitative analysis of verse and the study of sound-repetitions. The scaldic verse fully justifies the approach to word stress according to which "stress is not a force marking off a syllable and given a priori; rather, it is a mechanism referring syllables to one or another category" /17, p. 25/. But this statement is justified by the scaldic metrics only insofar as it operates with isolated words.

2.4. Quantity and stress. It should be noted in conclusion, that although scaldic verse is both genetically and functionally linked with the types of the epic short line (and scaldic devices are effective only against the background of epic poetry), the relations between the two prosodic features of the epic line - quantity and stress - appear turned upside down.

In AV quantity was subordinated to stress. The syllable length served as an additional means of marking the 'lifts' of the line. The role of quantitative rules is minimal in loose forms of AV (such as in LAY OF HILDEBRAND). held by emphatic phrasal stresses and thus justifying the name of tonic verse. The role of quantitative oppositions increases in the course of canonizing short-line schemes and/or reducing the range of syllable variation in the line. Their priority in relation to stress in scaldic verse is the natural consequence of its syllabism. But at the same time this is the result of the fact that the immediate prototype of the scaldic verse pattern is the short line, whose schemes are based on the prosodic structures of the word.

REFERENCES

/1/ M.L.West, "Indo-European Metre", in: "Glotta. Zeitschrift für griechische und lateinische Sprache", Bd.51, 1973, pp.161-187.
/2/ E. Sievers, "Die altgermanische Metrik", Halle (Saale), 1893.
/3/ J. Kuryłowicz, "Metrik und Sprachgeschichte", Wrocław-Warszawa-Kraków-Gdańsk, 1975.
/4/ V.M. Žirmunskij, "Vvedenie v metricu", in: V.M.Žirmunskij, "Teorija stixa", Leningrad, 1975.
/5/ Th.Cable, "The Meter and Melody of BEOWULF", Urbana, 1974.
/6/ S.J. Keyser, "Old English Prosody", in: "College English", v.3, N5, pp.331-356.
/7/ A. Heusler, "Deutsche Versgeschichte", Bd.I, T.I-II, Berlin, 1956.
/8/ M. Parry, "Studies in the Epic Technique of Oral Verse-Making, I: Homer and Homeric Style", in: "Harvard Studies in Classical Philology", v.43, 1932.
/9/ A.B. Lord, "The Singer of Tales", Cambridge, Mass., 1960.
/10/ A.Ch. Watts, "The Lyre and the Harp. A Comparative Reconsideration of Oral Tradition in Homer and Old English Epic Poetry", in: "Yale Studies in English", v. 169, New Haven, 1969.
/11/ "Old English Poetry. Fifteen Essays", ed. by R.P.Creed, Providence, 1967.
/12/ M.I. Steblin-Kamenskij, "Folklore and Literature in Iceland and the Problem of Literary Progress", in: "Scandinavica", Bd.II, N2, pp.127-136.
/13/ W.Ph.Lehmann, "The Development of Germanic Verse Form", Austin, Tex., 1956.
/14/ Th.A.Shippey, "Old English Verse", London, 1972.
/15/ P.G. Foote, "Beginnings and Endings: some Notes on the Study of Scaldic Poetry", in: "Les vikings et leur civilisation", Paris - La Haye (Mouton), 1976.
/16/ R. Frank, "Old Norse Court Poetry: The Dróttkvætt Stanza", in: "Islandica", Bd.42, Ithaca - London, 1978.
/17/ A. Liberman, "Germanic Accentology", v.I, "The Scandinavian Languages", Minneapolis, Minn., 1982.

КРИТЕРИИ ОРГАНИЗАЦИИ РИТМА ХУДОЖЕСТВЕННОЙ ПРОЗЫ, ВЕРЛИБРА И СИЛЛАБО-ТОНИЧЕСКОГО СТИХА

Л.В. Златоустова

Филологический факультет,
Московский гос. университет,
Москва 119899, СССР

ABSTRACT

The main prosodic characteristics of prose texts, vers libre and syllabo-tonic verse are defined on the basis of acoustic and perceptive analysis. A system of rhythm organization rules is suggested for the texts of various genres. An attempt is made to create formal criteria for distinguishing verse, vers libre and prose.

В последние два десятилетия проблема ритмической организации речи неоднократно привлекала внимание исследователей. Это вызвано и несомненным увеличением использования звучащей речи в сфере общения, и успехами в области возможностей исследования звучащей речи, и теми сведениями, которые дает нам современная наука о функциях ритма в живой природе и искусстве.

Вместе с тем нет достаточно полного описания единиц ритма, их структурной организации, недостаточно определены особенности ритма прозы, верлибра и стиха (силлабо-тонического).

В определениях ритма делается, как правило, акцент на неоднозначности метра и ритма, - это несомненно важно. В силлабо-тоническом стихосложении различают метр, понимаемый как условная схема расположения ударяемых и неударяемых слогов, а ритм - как фактическое распределение их.

Ритмическая организация стихотворного произведения связана с содержанием и со всей инструментовкой стихового произведения.

Однако в приведенном понимании ритма не учтен один из основополагающих фактов реализации единиц звучащей речи - факт объединения ударным слогом неударных слогов, что составляет фонетическое слово или ритмическую структуру (РС).

Термины синонимичны, они обозначают словоформу или сочетание словоформ, объединенных одним словесным ударением. Как правило, ритмическая структура равна знаменательному слову или сочетанию служебного и знаменательного слов. (Значительно реже, а для стихотворной речи крайне редко, в ритмическую структуру входят два знаменательных слова).

Аудитивные исследования ритмики фонетического слова не только подтвердили положение об организующей роли словесного ударения, но и показали возможность членения текста на ритмические структуры при подавлении речевого сигнала шумом, в результате чего смысл высказывания не опознавался, а его ритмическая структура сохранялась. (В качестве источника равномерного шума использовался генератор с полосой 20 - 20000 Гц).

При оптимальных условиях эксперимента - полный стиль произнесения, употребление наиболее частотных ритмических структур, соблюдение типичной их сочетаемости в рамках синтагмы, степень подготовленности аудиторов - возможна верная идентификация ритмических структур в 97% случаев. Этот процент подтверждает сведения, полученные путем инструментального и статистического анализа специфики организации предударной, ударной и заударной частей фонетических слов, а именно: сведения о степенях редукции гласных, типах консонантных стечений начал и концов слов, месте ударения и количестве слогов в слове и т.д.

Как показывают статистические исследования художественной прозы, верлибра и силлабо-тонического стиха, наиболее типичными ритмическими структурами (РС) для всех указанных типов текстов оказываются структуры односложные, двусложные - 2/I (цифровые обозначения ритмических структур построены так, что в числителе указано количество слогов, а в знаменателе - место ударения в структуре) и 2/2, трехсложные - 3/2, 3/3, четырехсложные - 4/3, пятисложные - 5/3. Однако проза, верлибр и художественная проза дают разное распределение частых ритмических структур. В стиховом тексте повышается процент односложных и двусложных структур, несколько сокращается процент трехсложных с ударением на первом слоге, понижается процент четырехсложных структур. Верлибр занимает некое промежуточное положение, т.е. в этом типе текста падает частота употребления многосложных структур.

Из сказанного следуют два положения, очень важные для дальнейшего изложения материала:

I. Ритмическая структура - объективно существующая в речи единица ритма, о чем свидетельствуют опыты со снятием лексического значения; типы структур существуют в памяти человека в некой обобщенной форме. Принципиально важными оказываются место ударения в структуре и набор предударных и заударных слогов с их специфической иерархией по степени редукции, наборами консонантов и вокальных элементов, позволяющих определить границы РС.

2. Ритмическая структура при конкретной речевой реализации имеет план содержания - лексико-семантическое значение словоформы или последовательности словоформ. Такой подход позволяет описать содержание и просодическую организацию в их единстве. Это особенно важно при анализе просодии опорных в эмоциональном или смысловом плане единиц речи. Сугубо формальное представление ритма текстов ограничивает эти возможности.

Рассмотрим характерные особенности ритма стихотворений А.Блока.

Равномерное чередование количества слогов в соседних строках - один из устойчивых признаков ритма лирики А.Блока. Эта особенность прослеживается во многих произведениях цикла "Стихов о Прекрасной Даме" ("Встану я в утро туманное", "За туманом, за лесами", "Тихо вечерние тени", "Душа молчит", "Я понял смысл твоих стремлений"). Аналогичное распределение количества слогов по строкам находим в лирике А.Пушкина, например, в стихотворениях "Предчувствие", "Стансы" и др. Более редкий случай - одинаковое количество слогов в строках всего стихотворения. В то же время количество ритмических структур, а следовательно, ударных слогов, не всегда одно и то же в каждой строке.

Вероятно, можно говорить о способах компенсаторной ритмизации, когда (например, в стихотворении "Ярким солнцем, синей далью...") меньшая регулярность повторяемости одинакового количества ритмических структур компенсируется равным количеством слогов в строке и повтором строк (см. табл. 2).

Рассмотренные параметры ритмической организации стихового текста, а именно: I) регулярное чередование количества слогов в соседних строках (9-8, 8-7, I0-9, 9-3); 2) регулярное чередование обязательной связи рифмы с данным типом РС; 3) регулярное чередование типов РС в начале соседних стиховых строк; 4) ограниченный набор РС, определяемый стиховым размером, достаточный для того, чтобы принять решение о типе текста – именно стихового, а не прозаического, даже в случае снятия графической подсказки – расположения текста по стиховым строкам.

Подтверждением этой точки зрения может служить эксперимент, при котором предлагалось прочитать без предварительной подготовки стиховые тексты (силлабо-тонические) с перекрестной рифмой; стихи были расположены как прозаический текст с сохранением знаков препинания.

Из 24 участников эксперимента, владеющих хорошей техникой чтения, 22 человека со второй строки начинали читать текст как стиховой, следовательно, ни рифма, ни строка не являлись для них подсказкой.

Закономерен вопрос: что же явилось критерием принятия решения о прочтении дикторами текста как стихового? Такие критерии легко сформулировать, исходя из анализа результатов инструментального эксперимента. В первую очередь это замедление темпа произнесения текста, которое связано в основном с вокализацией ритмических структур, проявляющейся в увеличении времени звучания гласных, преимущественно гласных ударного, первого предударного и конечного заударного слогов РС. Увеличение времени органически связано с другими просодическими особенностями стиха, которые также получены в результате акустического эксперимента. К ним относятся: а) однотипность повторяющихся РС; б) изменение регистра реализации частоты основного тона; в) специфическое изменение спектральной структуры гласных, заключающееся в расширении области первой и второй формант; г) появление устойчивой полосы усиления энергии в области четвертой форманты, где максимальная энергия реализуется на ударных гласных РС; д) специфика контура частоты основного тона (ЧОТ), характеризующаяся отсутствием контрастного выделения единиц текста (последнее не свойственно стиховой типологии набора единиц тональных контуров, отличительной чертой которых является суженность диапазона ЧОТ.

Предложенный перечень наиболее важных признаков, формирующих ритм стиха, есть способ выражения в звучащей речи синтагмы – единицы, представляющей собой семантико-синтаксическое единство, оформленное указанными выше фонетическими средствами. Наиболее простой случай – совпадение синтагмы и стиховой строки. Однако стиховая строка может включать две, реже три синтагмы:

(2 с.) Не сердись и прости / Ты цветешь одиноко./
(I с.) Да и мне не вернуть /
(2 с.) Этих слов золотых, / этой веры глубокой.../
(I с.) Безнадежен мой путь. /
(I с.) Неотвязный стоит на дороге,
(2 с.) Белый – / смотрит в морозную ночь./
(2 с.) Я – / на встречу в глубокой тревоге, /
(3 с.) Он, / шатаясь, / сторонится прочь /

Независимо от количества синтагм, стиховая строка сохраняет ритмическое единство благодаря внутренней структурированности единиц РС и синтагм, составляющих строку. Определенная роль здесь принадлежит рифме.

В организации ритма графическая урегулированность текста не играет ведущей роли – она лишь способствует выявлению его при чтении, выполняя ту же функцию, какую выполняет пробел между словами в любом графическом представлении речи.

Значит ли в таком случае, что для стихов ранней лирики А.Блока строка – формальный показатель графики, не несущий ритмообразующей функции? Да, значит, если остановиться в анализе ритма только на тех параметрах (хотя они и занимают ведущее место в ритмообразовании), которые здесь рассмотрены. Однако ритмообразующей функцией в звучащей речи обладает пауза, которая подчеркивает ритмическую схему синтагмы, способствует выявлению ритма на уровне строфы и целого стихотворения, выявлению регулярной повторяемости хорошо структурированных отрезков речи.

В табл. I и 2 представлены значения в мс распределения пауз в текстах.

Таблица I.

Стихотворение "Милый друг"

Кол-во слогов	Длительность (в мс)		
	стиха	паузы межстиховой	внутристихов.
I0	2838	552	572
3	I304	8I6	-
I0	3I48	94	652
3	867	528	-
I0	3264	-	звуч. п.
3	I042	9I4	-
I0	3488	496	II2
3	982	830	-
I0	37I8	660	I62
3	I446	506	-
I0	3876	II0	I88
3	I570	-	-

Даже беглая оценка приведенных значений говорит о ряде особенностей употребления пауз. Паузой завершается каждая (за редким исключением) строка, на границе строки пауза больше, чем внутри строки в случаях ее синтагматического членения.

Таблица 2.

Кол-во слогов	Длительность (в мс)		
	стиха	паузы межстиховой	внутристиховой
8	2238	482	звуч. п.
8	2696	462	-
8	2206	2I8	-
8	2888	9I2	-
8	2898	4I4	I98
8	2524	562	-
8	2594	4I0	-
8	2606	-	-

Распределение пауз теснейшим образом связано с мелодическим оформлением текста, которое выполняет для стиха прежде всего функцию создания тональной рамки регулярно чередующихся структурированных единств, заданных последовательностью типов РС и типов синтагм, составляющих строку. Собственная ритмизирующая функция стиховой строки заключается во взаимодействии паузы и мелодического контура.

Сравнение текстов верлибра (Тургенев И., Куприянов В.) со стиховыми текстами показывает значительно большее разнообразие типов РС в верлибре. Однако ряды повторяющихся РС, составляющих "блоки" текста верлибра, не типичны для художественной прозы. Например, у И.Тургенева находим следующее распределение типов РС в последовательности синтагм:

Теперь зима; / мороз запушил стекла окон/
2/2 2/2 2/2 3/3 2/I 2/I
в темной комнате / горит одна свеча //
2/I 3/I 2/2 2/2 2/2

Следствие подобной ритмической организации – вся просодическая инструментовка текста: повторяемость РС, замедление темпа, сужение диапазонов ЧОТ, выровненность ЧОТ, создание типичной тональной рамки синтагмы, близкой к характеристикам стиховой, сокращение длительности синтагмы по количеству РС. Сходство просодического рисунка в синтагмах верлибра и стиха определяется общим художественным заданием текстов, где основная функция просодической инструментовки не создание просодических контрастов, свойственных прозе, а наоборот, стремление к снижению их.

В значительном количестве произведений стиха и верлибра социальное сознание писателя как основной предмет художественного воспроизведения жизни направлено "или в глубь самого себя или на явления внешнего мира, на социальные и личные отношения бытия, на явления природы" /I/. Сказанное находит свое выражение в сфере просодии стиха и верлибра.

Основным отличием стихотворного текста от верлибра является отсутствие в нем некоторых типов ритмического параллелизма, а именно: закономерных повторов ритмо-мелодических схем синтагм (РМСС), строк, рифм. В верлибре редки случаи полного совпадения в синтагмах тональных рисунков, что составляет специфику силлабо-тонического стиха, где одинаковый тип тональной рамки строки есть существенный компонент ритмики стиха. Верлибр и по параметру интенсивности характеризуется большей неоднородностью оформления синтагм, что создает градации, стиху не свойственные. Для интенсивности внутри синтагм верлибра характерно появление контрастных выделений отдельных единиц, что сближает верлибр по данному показателю с прозаическим текстом /2/.

В отличие от силлабо-тонического стиха и верлибра прозаический художественный текст направлен прежде всего на выполнение коммуникативной функции с подчинением ей функции членения текста, выражение модально-эмоциональных отношений с учетом экстралингвистических факторов, способствующих созданию художественной реальности.

Вместе с тем основой просодии прозы также является совокупность средств ритма. Это: типы РС и их последовательность в синтагме, синтагма и ее объем в количестве РС. Однако в прозе упорядоченность фразово-синтагматической организации имеет собственные закономерности, включающие типичные для каждого языка повторяемость синтагм по количеству в них РС, фразовое и фоноабзацное разбиение.

Нельзя не остановиться на специфике реализации типов фразовых акцентов в текстах разных жанров. Так, функция синтагматического фразового акцента – объединять просодическими средствами последовательность РС, что создает РМСС. Однако типология РМСС различна для трех рассматриваемых жанров.

Фразовые выделительные акценты также специфичны в своей реализации по жанрам. Если в прозаическом тексте они выражены контрастно средствами частоты основного тона и сегментными спектральными характеристиками, то в стихотворном – отсутствуют резко-контрастные выделения ЧОТ и интенсивности, а сегментные спектральные показатели в силу особенностей общей просодической организации стихового текста – создание "стиховой тесситуры" – имеют нестандартные показатели спектра, особенно слогов и слов, стоящих под фразовым акцентом.

На основании полученных экспериментальных данных целесообразно выделить для стиховых текстов и верлибра особый выделительный тип ударения – эмоционально-поэтический, в отличие от логического и эмфатических.

На основании всего сказанного выше можно предложить следующую схему просодии трех типов текстов. Естественно, что приведенные количественные данные о типах РМСС отражают исследованный материал.

Рис. 1

На рис. I дана обобщенная схема, показывающая роль ритма в образовании просодии разных жанров текстов.

ЛИТЕРАТУРА

/I/ Поспелов Г.Н. Лирика. – М., изд. МГУ, 1976, с. 63.
/2/ Кедрова Г.Е. Фоностилистические варианты оформления текста. – Автореф. канд. дисс. М., 1985.

ACADEMY OF SCIENCES OF THE ESTONIAN S.S.R.

INSTITUTE OF LANGUAGE AND LITERATURE

CONTENTS

INDEX OF AUTHORS